UNIVERSITÉ DE GRENOBLE

Mémoire

pour obtenir le diplôme d'

HABILITATION À DIRIGER DES RECHERCHES

Présenté par

Sabine Crépé-Renaudin

préparé au **Laboratoire de Physique Subatomique et de Cosmologie**

Outils et analyse en physique des particules : morceaux choisis

La grille de calcul et de stockage pour le LHC : de la mise en place d'un nœud de grille à l'utilisation de la grille par l'expérience ATLAS

Mesure de la section efficace de production top-antitop avec l'expérience DØ auprès du Tevatron

HDR soutenue publiquement le 15 juillet 2013
devant le jury composé de :

Christophe Furget
Professeur, Président
Stéphane Jézéquel
Directeur de recherche, Rapporteur
Dominique Pallin
Directeur de recherche, Rapporteur
David Rousseau
Directeur de recherche, Rapporteur
Laurent Vacavant
Directeur de recherche, Examinateur

Table des matières

Préambule

Ce document marque une étape dans mon parcours de physicienne et il m'a semblé important d'y décrire les différentes facettes de mes activités de recherche. Il n'est pas possible d'être exhaustif ici bien sûr, cependant, j'ai délibérément choisi de ne pas me restreindre à la partie de mes activités qui concerne l'analyse de données mais de décrire aussi deux autres aspects de mon travail, à savoir une partie plus technique qui concerne la grille de calcul ainsi qu'un court volet sur la communication au grand public. Ces deux derniers sujets sont en effet importants à mes yeux bien qu'en général moins valorisés que l'analyse. Bien sûr, sans l'implication de physiciens dans la construction et le fonctionnement des détecteurs ou du calcul, l'étape d'analyse ne pourrait pas être envisagée. Quant à la communication au grand public, elle représente en quelque sorte l'aboutissement du travail de recherche car elle permet le retour, le partage avec la société des nouvelles découvertes (qu'elle a d'ailleurs financées) mais aussi du travail dans nos laboratoires.

On notera qu'au vu de la complexité croissante des expériences de physique des particules, il devient difficile de ne pas se spécialiser dans l'une de ces activités pour gagner en efficacité et capitaliser une certaine expertise. J'ai pourtant choisi jusque-là de garder ce que je pense être un bon équilibre entre ces activités et décidé de le retranscrire dans ce document. Ce mémoire ne retracera pas toutes les activités de recherche que j'ai menée depuis la fin de ma thèse et encore moins celles qui les accompagnent (vie du laboratoire, administration et autres) qui n'entrent pas dans le cadre de ce document. J'ai donc sélectionné des sujets parmi mes plus récentes activités dans les deux dernières expériences auxquelles j'ai participé : DØ auprès du Tevatron à Fermilab et ATLAS auprès du LHC au CERN.

Introduction

Comme je l'ai précisé dans le préambule, afin que ce mémoire soit le reflet de mes activités de recherche, j'ai choisi de rendre compte de différentes facettes de mon travail.

Ce document se partagera donc en quatre parties :

- La première, en guise d'introduction, décrira les différentes étapes nécessaires à la recherche expérimentale en physique des particules depuis la conception des expériences jusqu'à la publication des résultats.
- La deuxième partie, technique, introduira la grille de calcul qui permet entre autres aux physiciens d'ATLAS de traiter partout dans le monde les données de cette expérience ; cette partie décrira aussi le nœud de grille du LPSC dont je suis la responsable scientifique.
- La troisième partie relatera un travail d'analyse des données que j'ai effectué alors que j'étais membre de la collaboration DØ. Il s'agit de la mesure de la section efficace de production de paires de quarks top dans le canal lepton+jets.
- En guise de conclusion sur la partie analyse de données, je décrirai ensuite mon projet de recherche au sein de la collaboration ATLAS.
- Enfin, je décrirai très brièvement les activités liées à la communication grand public que je mène au laboratoire depuis plusieurs années.

Première partie

Les différentes étapes nécessaires à la recherche expérimentale en physique des particules

Pour répondre à nombre des questions que se posent les physiciens des particules, il s'est avéré nécessaire de concevoir des expériences de grande complexité et de grande envergure : les accélérateurs et les détecteurs de particules opérés par des collaborations internationales qui se partagent le travail ainsi que le financement. À toutes les étapes de la vie d'une expérience, de la conception à l'opération, s'entremêlent les études théoriques et d'analyse et les tâches techniques liés aux détecteurs et au traitement des données, sans compter les tâches administratives et la recherche de financement. Dans cette partie sont résumées les différentes phases d'un tel projet en prenant l'exemple du plus récent d'entre eux : le LHC et ses détecteurs.

Dans un premier temps, les accélérateurs, détecteurs et logiciels sont conçus en fonction des performances déterminées par des simulations informatiques selon des hypothèses théoriques à tester. Les contraintes techniques et financières prises en compte, le travail sur des techniques innovantes, la recherche et développement, les tests de prototypes permettent la conception de ces appareils. Pour le LHC et ses détecteurs, la grande majorité de ce travail a été effectuée dans des laboratoires de recherche répartis dans le monde entier.

Une fois le projet défini, la phase de construction commence. La taille des détecteurs du LHC a rendu nécessaire le passage à l'industrie pour la production en grande quantité de leurs composants. Ceux-ci sont ensuite testés et assemblés dans les laboratoires responsables ou au CERN. L'installation et l'assemblage final ont lieu au CERN. En parallèle, les infrastructures informatiques sont construites et les logiciels nécessaires au traitement et à l'analyse des données conçus.

Aux premiers tests en vraie grandeur, succèdent la mise en route de l'accélérateur et des détecteurs. La préparation, les tests et les prototypes en amont facilitent le démarrage et la compréhension des détecteurs. Ceux-ci doivent être étalonnés et testés en vraies conditions.

La prise de données commence alors. Les performances des détecteurs doivent être mesurées, optimisées et adaptées sans cesse aux modifications des conditions expérimentales. Les données doivent être traitées, distribuées et analysées. Enfin les études doivent être validées par la collaboration et enfin publiées. Les résultats et la façon dont on les a obtenus peuvent alors être expliqués et diffusés au grand public.

Les études et outils décrits dans ce document sont relatifs à la période de prise de données et à l'analyse de ces dernières. Pour préciser le cadre de ce travail, les différentes étapes de l'analyse sont donc exposées dans les paragraphes qui suivent.

Il y a deux grandes catégories d'analyses :

– les mesures de caractéristiques de particules ou de phénomènes déjà mis en évidence,
– les recherches de particules ou de phénomènes nouveaux.

Dans les deux cas, il faut sélectionner les collisions ou évènements d'intérêt pour l'étude et

évaluer la proportion de ces évènements, communément appelés signal, et celle des bruits de fond c'est à dire des évènements provenant de phénomènes physiques différents de ceux qu'on veut étudier mais avec une signature proche voire similaire à celle du signal.

S'il s'agit d'une mesure des caractéristiques d'un phénomène connu, il faut déterminer la (ou les) variable(s) expérimentale(s) qui permettront d'obtenir cette mesure à partir d'un échantillon, en général de grande pureté, et évaluer les erreurs liées à la mesure. Ces incertitudes sont la somme de l'incertitude statistique et des systématiques liées à notre méconnaissance ou imprécision théorique et expérimentale ainsi que de la pollution induite par le bruit de fond restant dans l'échantillon.

S'il s'agit de mettre en évidence un phénomène nouveau, les données sélectionnées sont comparées à nos connaissances théoriques ou expérimentales les mieux validées. Des outils statistiques sont utilisés pour mettre en évidence des éventuelles déviations dans les données, statistiquement incompatibles avec le modèle de physique établi.

Dans tous les cas, les données réelles issues des collisions doivent être confrontées à nos connaissances les mieux établies pour s'assurer d'une part de notre bonne compréhension du fonctionnement et des réponses des détecteurs et d'autre part, pour évaluer les bruits de fond d'une analyse donnée ou encore pour mettre en évidence de nouveaux phénomènes. Pour ce faire, on simule entièrement des collisions de façon informatique. Celles-ci sont produites à l'aide de générateurs Monte Carlo qui génèrent des particules selon les probabilités de la théorie étudiée. Le passage des particules dans les détecteurs est simulé grâce à l'aide d'un logiciel appelé Géant [1] dans lequel les détecteurs sont décrits avec précision. En sortie de simulation les évènements sont similaires à ceux produits par les collisions réelles. Collisions simulées et collisions réelles sont ensuite reconstruites par le même programme qui, à partir des signaux des détecteurs et de l'étalonnage de ces derniers, permet de reconstituer les particules issues de la collision.

Pour toutes ces étapes, des ressources informatiques sont nécessaires. La deuxième partie de ce document décrit la solution choisie par les expériences LHC pour stocker et traiter leurs données : la grille de calcul. La grille est composée de différents sites répartis tout autour du monde. Le site du LPSC sera décrit dans cette partie. Dans la troisième partie, une analyse des données de l'expérience DØ auprès du Tevatron sera décrite. Elle a pour but la mesure de la section efficace de production top-antitop et illustre la partie analyse de données dans ce document.

Deuxième partie

La grille de calcul pour le LHC et le nœud de grille du LPSC

Le « Large Hadron Collider » (LHC) du CERN a désormais pris le relais du Tevatron. En effet, depuis mars 2010 et les premières collisions de protons à 7 TeV, le LHC a repoussé la frontière en énergie du domaine d'exploration accessible aux expériences de physique des particules. L'énergie des collisions n'est cependant pas la seule limite que le LHC a dépassée. Ce collisionneur fournit et fournira une quantité et un flux de collisions que les expériences devront enregistrer et analyser bien au delà de ce que ses prédécesseurs ont réalisé. Pour la seule expérience ATLAS, environ 100 millions de canaux électroniques devront être lus à chaque collision. Les paquets de protons du LHC entrent actuellement en collision toutes les 50 ns, cet intervalle de temps devrait passer dans les années qui viennent à 25 ns ce qui devrait amener le taux moyen de collision à 30 MHz [1].

Si la fréquence des collisions est si grande, c'est parce que celles qui produisent des évènements intéressants sont extrêmement rares, de l'ordre de 1 sur plusieurs centaines à plusieurs milliers de milliards. De plus, comme il n'est pas possible d'enregistrer toutes les collisions, le système de déclenchement d'ATLAS permet de sélectionner les plus intéressantes. Ainsi, le taux d'évènements réellement enregistré était de plus de 300 Hz en 2011 et a dépassé les 500 Hz en 2012. En sortie du détecteur, les données pour une collision correspondent à une taille moyenne de 1,6 Mo, ce qui revient à écrire un flux de données de l'ordre de 600 Mo/s. L'expérience ATLAS, tout comme ses voisines, devra donc stocker et analyser plusieurs Péta-octets [2] de données chaque année. Il faut quelques milliards d'heures CPU (600 ans) pour traiter un Péta-Octet de données et les données traitées devront être accessibles au millier de physiciens qui travaillent à leur analyse.

Il est apparu très vite qu'une institution comme le CERN ne pouvait ni accueillir ni financer l'infrastructure de calcul nécessaire au traitement et au stockage des données du LHC. Les expériences LHC se sont donc tournées vers le calcul distribué qui permet d'exploiter de façon optimale les ressources informatiques quelle que soit leur localisation. Il a fallu cependant adapter le concept de grille alors surtout utilisé pour des tâches nécessitant une grande puissance de calcul pour y intégrer aussi l'aspect stockage et distribution d'une grande quantité de données. Le tout devant être opérable en toute sécurité et accessible quelle que soit la localisation des membres des expériences LHC.

Je décrirai dans les paragraphes qui suivent le principe et le fonctionnement de la grille de calcul LHC en allant du schéma général jusqu'à la description d'un site particulier (celui hébergé par le LPSC) en passant par l'organisation nécessaire à l'échelle des expériences. Étant donné ma participation à l'expérience ATLAS, ces éléments seront souvent décrits via le prisme des choix et du fonctionnement de cette expérience LHC particulière. Par ailleurs, on verra que le modèle d'utilisation de la grille de calcul par les expériences LHC a évolué pendant la première phase de prise de données (Run 1). Je décrirai donc les schémas d'or-

1. Une collision toutes les 25 ns correspond à un taux de collision de 40 MHz mais des intervalles plus grands sont nécessaires parfois pour permettre par exemple l'injection des protons dans le LHC. Le taux moyen de collision est égal au nombre total de paquets multiplié par le nombre de tours par seconde dans le LHC soit $2808 \times 11\,245 = 31.6$ MHz.

2. ou Po, soit 10^{15} octets.

ganisation tels qu'ils ont été conçus au démarrage du LHC (2009-2010) et je soulignerai les évolutions du modèle jusqu'à aujourd'hui, à la fin du Run 1.

1 La grille de calcul : principes généraux et application au calcul LHC et à l'expérience ATLAS

1.1 Principe

Une grille de calcul et, devrait-on ajouter, de stockage, repose sur un ensemble d'infrastructures informatiques appelés « sites » de localisation géographique quelconque. En préalable, chacun de ces sites doit être connecté au réseau internet. D'autre part, chacune de ces infrastructures propose à des communautés d'utilisateurs des services tels que des espaces de stockage et de la puissance de calcul avec des interfaces communes :

- Les « Computing Elements » (CE) servent d'interface avec le système qui distribue les tâches de calcul aux processeurs (« batch system ») et avec les logiciels nécessaires aux utilisateurs.
- Les « Storage Element » (SE) représentent l'interface aux systèmes de stockage.

Le nœud central qui permet de faire fonctionner la grille est un ensemble de logiciels appelé intergiciel (« middleware » en anglais). Ce dernier fait l'interface entre les demandes des utilisateurs et les composants des différents sites.

Un autre aspect important lié à une infrastructure de grille est la présence d'un système d'authentification et d'autorisation permettant aux utilisateurs d'accéder aux ressources des sites qui ont accepté de travailler pour la communauté à laquelle ils appartiennent. Les différentes communautés sont identifiées via ce qu'on appelle des « Virtual Organisations » ou VO. Pour appartenir à une VO, un utilisateur doit au préalable être identifié par une autorité de certification reconnue par l'ensemble des acteurs de la grille.

D'autre part, chaque site publie ses propriétés sur un système commun d'information, et les ressources utilisées en fonction des communautés d'utilisateurs doivent pouvoir être comptabilisées.

Enfin, pour permettre l'accès sécurisé aux données quelle que soit leur localisation et uniquement pour les utilisateurs autorisés, les fichiers sont identifiés sur la grille grâce à un identifiant unique, le GUID (« Globally Unique Identifier »), associé à un nom logique plus lisible (« Logical File Name » ou LFN). Une partie du LFN permet d'identifier la communauté ou VO auquel le fichier appartient. Cet identifiant est associé aux adresses physiques du fichier appelée PFN « Physical File Name » via le service RMS « Replica Manager Service » [3]. En général, les communautés d'utilisateurs associent le LFN à des « méta-données » (des informations permettant de caractériser les données) qui permettent la recherche des fichiers dans une base de données qui leur est propre.

3. Les adresses physiques du fichier peuvent aussi prendre la forme d'un SURL « Site URL » lorsque le site utilise SRM [2] comme interface pour la gestion de son stockage.

1.2 La grille de calcul LHC

Au LHC, la grille de calcul doit permettre d'une part, de sauvegarder les données produites par les expériences et d'autre part, de les traiter en plusieurs étapes et à différents niveaux, de façon centralisée ou par n'importe lequel des membres des expériences.

La grille de calcul LHC, s'est construite autour de trois grilles principales : EGEE devenue EGI [3] présente en particulier en Europe, OSG [4] aux États-Unis, ainsi que NorduGrid [5], grille régionale des pays du nord de l'Europe à l'origine, qui possèdent chacune leur propre intergiciel. À chaque expérience LHC correspond une VO particulière. Les expériences assurent à leurs utilisateurs l'accès à leur VO et contrôlent cet accès ce qui garantit l'authentification des utilisateurs pour les sites. Les sites sont libres de choisir les VO qu'ils accueillent et leur allouent une partie des ressources qu'ils possèdent. En général, ils soutiennent les VO liées directement aux activités de recherche des laboratoires ou instituts auxquels ils sont attachés.

La grille LHC (WLCG : Worldwide LHC Computing Grid [6]) est actuellement composée de 170 sites répartis dans 36 pays qui permettent l'accès aux données aux quelques 8000 physiciens qui travaillent sur les expériences LHC. En 2012, l'ensemble de ces sites ont mis à disposition du LHC [7] :

- 1,8 millions de HEP-SPEC06 [4] de puissance de calcul, ce qui représente environ 300 000 processeurs,
- 175 Po d'espace disque,
- 170 Po d'espace sur bandes.

Les contributions des sites et donc de leur agence de financement sont discutées chaque année selon un protocole d'accord (MOU : Memorandum Of Understanding) avec WLCG [8].

Les sites sont classés en 4 niveaux appelés « Tier » en anglais :

- **Le Tier 0**
  Il s'agit du centre de calcul du CERN. Son rôle est d'une part de sauvegarder sur bande, au fur et à mesure de leur arrivée, les données produites par les expériences en mettant à jour le catalogue qui permet de les référencer. Ces données sont immédiatement répliquées dans un des sites de niveau 1 afin qu'il existe toujours, pour des raisons de sûreté, deux copies des données brutes délivrées par les expériences à deux endroits différents. D'autre part, il effectue le premier traitement des données des expériences. Le Tier 0 du CERN, noté T0 par la suite, a offert en 2012 environ 30 Po d'espace disque, 70 Po sur bande et environ 65 000 cœurs (350 000 HEP-SPEC06) à l'ensemble des expériences LHC.

- **Les Tiers 1**
  Les sites de niveau 1, notés T1s, sont de grands sites nationaux. Ils sont au nombre de 11 et sont situés au Canada, en France, en Allemagne, en Italie, aux Pays-Bas, dans les pays nordiques (T1 distribué), en Espagne, à Taïwan, au Royaume-Unis et il y a 2 T1s aux États-Unis. Ils ont fourni à la communauté LHC en 2012 environ 600 000 HEP-SPEC06, 65 Po d'espace disque et 100 Po d'espace sur bande. Ces sites se partagent

4. Le HEP-SPEC06 est une unité de référence utilisée en physique des hautes énergies permettant d'évaluer la puissance de calcul d'un processeur indépendamment de ses caractéristiques techniques ; pour donner une idée, une machine de 2010 avec 8 cœurs fait environ 100 HEP-SPEC06.

une copie des données brutes en général sur bande afin de sauvegarder les données au CERN. Pour ce faire, ils sont reliés au T0 par un réseau privé LHCOPN [9]. Ils permettent aussi de traiter les données brutes, de stocker des données sur disque, d'effectuer des simulations et pour certains sont ouverts à l'analyse par les membres des expériences LHC.

- **Les Tiers 2**
  Les sites de niveau 2 (T2s) ont des contacts privilégiés avec un des T1s. Ils permettent de produire les données de simulation et accueillent les tâches d'analyse des utilisateurs. Ils hébergent donc aussi de façon partagée les données des détecteurs ou de simulation une fois traitées. Il y a environ 140 sites de type T2. Comme le T0 et les T1s, ces sites s'engagent chaque année auprès des expériences LHC à fournir des ressources de stockage et de calcul. En 2012, les T2s ont mis à disposition des expériences LHC 80 Po de disque et 800 000 HEP-SPEC06 pour le calcul.
- **Les Tiers 3**
  Enfin les sites de niveau 3 (T3s) sont des sites qui n'ont pris aucun engagement en terme de ressources minimales à fournir aux expériences LHC. Ils mettent néanmoins à disposition leurs ressources, en particulier pour l'analyse, en priorité à leurs utilisateurs locaux. Ces sites peuvent aussi fournir des ressources pour effectuer de la simulation. Souvent les sites T2s possèdent aussi une partie T3.

1.3 La grille de calcul pour l'expérience ATLAS

Chaque expérience LHC a son propre modèle de calcul et ses outils (parfois partagés avec une ou plusieurs autres expériences) pour traiter ses données sur la grille. Les paragraphes qui suivent décrivent uniquement le fonctionnement de l'expérience ATLAS.

1.3.1 Topologie

Outre le T0 du CERN, l'expérience ATLAS peut compter sur environ 130 sites qui acceptent la VO ATLAS dont 10 sites de niveau 1 et 70 sites de niveau 2. Cela représentait en 2012 :

- pour le T0 : 111 000 HEP-SPEC06 (6000 tâches peuvent être traitées en parallèle), 9 Po d'espace disque et 18 Po de bandes ;
- pour les T1s : 285 000 HEP-SPEC06, 30 Po d'espace disque et 38 Po de bandes ;
- pour les T2s : 328 000 HEP-SPEC06 et 45 Po d'espace disque.

Les sites T2s sont regroupés dans ce qu'on appelle un « nuage » autour d'un T1 avec lequel ils ont des relations privilégiées, en particulier une bonne connexion réseau. En général les sites T2s sont liés au T1 de leur pays quand il existe ou à un T1 proche géographiquement mais ce n'est pas toujours le cas. Le nuage français par exemple, regroupe le T1 français situé au CC-IN2P3 ainsi que tous les sites des laboratoires français qui participent à ATLAS mais aussi les sites roumains, chinois et japonais. Le modèle de calcul d'ATLAS, tel que conçu avant le démarrage de la prise de données en 2009 [10], était basé en grande partie sur cette hiérarchie. Comme on le verra, avec l'expérience, l'évolution des besoins et des performances techniques notamment en terme de réseau, cette organisation a évolué vers

un système beaucoup plus souple.

1.3.2 Les étapes de traitement des données : des données brutes aux résultats de physique

Afin d'analyser les données récoltées par les détecteurs de l'expérience ATLAS et produire des résultats de physique, plusieurs étapes sont nécessaires. En voici une description succincte :

- **Référencement et sauvegarde des données**
  Les détecteurs produisent des données brutes numérisées (raw data) qui doivent avant tout être référencées et sauvegardées. Cette étape se fait dans un premier temps au CERN, au niveau du T0 [11], où les données sont regroupées en fichiers, enregistrées sur bande et référencées dans des bases de données sous Oracle. Les données brutes sont ensuite réparties dans les T1s (environ 1/10 ème des données pour chacun des 10 T1s) pour avoir une sauvegarde de l'ensemble à l'extérieur du CERN.

- **Premier traitement des données**
  Les données brutes issues des détecteurs doivent être traitées, « reconstruites », pour pouvoir être ensuite analysées. Cette étape est aussi effectuée par le T0 du CERN en moins de 48h. Cette mise en forme des données nécessite cependant la connaissance précise de l'état et de l'étalonnage des détecteurs au moment de la prise de données correspondante. Dans ce but, les données sont subdivisées en fonction du temps :
  - en « runs » ce qui correspond au temps de vie des faisceaux dans le LHC (jusqu'à une quinzaine d'heures) et à une configuration stable de la prise de données ;
  - en « blocs de luminosité » d'une ou deux minutes pendant lesquelles les conditions de luminosité des faisceaux peuvent être considérées comme stables.

  Les données sont aussi réparties en différents flots de données (« stream ») non exclusifs en fonction de leur classement par le système de déclenchement. Une petite partie des données (environ 10 Hz), prélevée dans chaque catégorie de déclenchement [12], est classée dans le flot noté « express stream » pour être traitée sans délai (moins de 8 heures) afin de vérifier rapidement la bonne qualité des données et donc le bon fonctionnement des détecteurs. Ces données, avec celles issues du « calibration stream », également traitées dès leur arrivée, permettent de mesurer l'alignement et l'étalonnage des différents détecteurs. Les résultats de ces études d'alignement et d'étalonnage ainsi que les conditions de prise de données sont enregistrés dans des bases de données. On notera que certaines données d'étalonnage sont traitées dans quelques centres T2s. Les streams de physique qui contiennent l'ensemble des données à analyser sont ensuite reconstruits dans un deuxième temps en utilisant ces résultats.

  L'ensemble de ces étapes impose un délai d'environ 48h entre la prise de données et leur reconstruction. Le T0 reconstruit les données sous différents formats (ESD, AOD, etc. : voir section 1.3.3), les fichiers de sortie sont ensuite enregistrés et distribués aux T1s et aux T2s via le système de gestion des données d'ATLAS.

- **Production Monte Carlo**
  Comme il a été développé dans la première partie de ce document, les avancées en

physique des particules se font en confrontant les données issues des collisions du LHC à celles prédites par les modèles théoriques après simulation des détecteurs. La génération d'évènements Monte Carlo reflétant les théories étudiées, leur traitement par les simulations des détecteurs puis l'étape de reconstruction qui est identique à celle des données réelles se font principalement dans les centres T1s et T2s de la grille de calcul. Les données produites sont ensuite référencées et stockées dans le centre T1 dont le T2 dépend. L'ensemble de ces étapes est regroupé sous le terme de « production Monte Carlo ».

On notera que la gestion de la production a été en grande partie automatisée, ce qui permet d'alléger de façon significative le travail consistant à vérifier que toutes les tâches se sont correctement terminées et à les soumettre à nouveau dans le cas contraire.

– **Analyse**
Enfin, les données réelles et de simulation sont analysées par les physiciens d'ATLAS. Ce travail se fait en plusieurs étapes.

Des études fines sur les détecteurs et les logiciels et techniques de reconstruction permettent d'améliorer la qualité des objets reconstruits. Ce travail se fait en continu, l'accumulation des données réelles permettant toujours de mieux comprendre et maîtriser la réponse des détecteurs et ainsi d'affiner la qualité des objets ou particules issus du traitement informatique. Pour bénéficier de ces améliorations, l'ensemble des données sont reconstruites une à deux fois par an avec une nouvelle version du logiciel d'ATLAS prenant en compte ces avancées. Cette reconstruction se fait à partir des données brutes et donc principalement dans les centres T1s. Certains centres T2s ont cependant été utilisés aussi dans la dernière campagne de 2012.

Les données sont ensuite triées et sélectionnées en fonction des études de physique envisagées pour permettre d'alléger au maximum la quantité de données à traiter dans la phase finale d'analyse. Ce travail est en général géré au niveau des groupes de physique et les données obtenues sont la base des analyses de ce groupe. Cette étape s'effectue dans les sites T2s et T3s et bénéficie maintenant des outils de la production centralisée.

L'étape d'analyse finale se fait ensuite avec les ordinateurs personnels des physiciens, des fermes de calcul hors grille ou les T2s et T3s en fonction de la quantité de données à traiter et de la complexité des calculs nécessaires.

1.3.3 Les outils de la grille ATLAS

Différents outils ont été développés pour permettre de gérer les différentes tâches liées au traitement distribué des données. Ils peuvent être classés en trois grandes catégories :

– **La gestion des données**
La gestion des données dans ATLAS est assuré par DDM « Distributed Data Management system » [13]. Ce système doit permettre de cataloguer l'ensemble des données, de les transférer vers les sites ou de les supprimer des sites en respectant la politique de distribution déterminée par la collaboration. L'actuelle implémentation logicielle de DDM est appelée DQ2 « Don Quixote 2 » [14], une nouvelle implémentation, Ru-

cio [15], est en cours de développement.

Le premier élément de la gestion des données d'ATLAS est le fichier qui contient une collection d'évènements (de collisions) sous forme d'une série d'objets C++. En général, les fichiers peuvent être regroupés en catégories d'intérêt commun pour les utilisateurs en fonction du type de données qu'ils contiennent et de la façon dont ils ont été traités. En fonction de ces catégories, les fichiers sont donc regroupés en datasets. Ces datasets sont non seulement l'élément de base utilisé pour la réplication des données par DDM pour les transferts entre sites mais aussi pour le traitement des données qui s'en suit. Un fichier peut appartenir à plusieurs datasets. Les datasets peuvent être ouverts (on peut leur ajouter d'autres fichiers) ou fermés (aucun fichier ne peut être ajouté). Avec l'accumulation des données du LHC, il est apparu nécessaire de créer un structure supplémentaire, le container, qui regroupe un ensemble de datasets. Ces containers permettent de sélectionner un ensemble de datasets de même type mais qui sont en général trop gros pour être manipulé simplement par les outils de l'expérience.

Les fichiers contiennent indifféremment différents formats de données :

- RAW : il s'agit des données (binaires) telles qu'elles sont produites par l'acquisition ; dans ce format un évènement occupe en moyenne 1,6 Mo.
- ESD (Event Summary Data) : il s'agit du format de données après reconstruction complète des données RAW ; ce format comprend les informations sur les hits des détecteurs de traces ainsi que les traces reconstruites, les cellules et les amas (clusters) reconstruits du calorimètre ainsi que les objets combinés issus de la reconstruction ; le tout est sous la forme d'objet POOL/ROOT [16] ; chaque évènement occupe en moyenne 1 Mo.
- AOD (Analysis Object Data) : il s'agit d'un résumé des objets reconstruits (toujours au format POOL/ROOT) avec des informations directes sur les objets physiques (électrons, muons, jets ...). Chaque évènement dans ce format occupe environ 200 ko.
- DPD (Derived Physics Data) : il s'agit du format utilisé comme base pour la plupart des analyses et il est optimisé en fonction de ces dernières après filtrage des évènements et nettoyage des variables non utiles ; cela permet de diviser par 10 environ la taille des évènements AOD mais ce gain dépend beaucoup des analyses considérées.
- TAG : il s'agit d'une base de données d'évènements ou de fichiers ROOT, permettant d'utiliser rapidement des évènements sélectionnés dans les fichiers d'AOD ou d'ESD.

Fonctionnement de DQ2 : DQ2 doit être capable d'enregistrer et de placer les quelques milliers de datasets produits chaque jour soit par l'acquisition des données soit par la production Monte Carlo. Au centre de ce système se trouve donc une base de données Oracle ou Catalogue Central qui référencie les fichiers (avec leur GUID, leur LFN, leur taille ...), les datasets (avec leur nom, leur propriétaire, leur date de création, leurs versions, ...) ainsi que leurs répliques, et enfin les demandes de réplication des utilisateurs. Les points d'accès à DQ2 sont des serveurs web Apache. L'utilisateur interagit principalement avec les catalogues via des outils spécifiques ce qui lui permet de définir ou rechercher des datasets ou de demander leur réplication sur un site. Les « Site Services » gèrent les transferts tels que définis dans le catalogue et envoie leur état au service de monitoring. Si des fichiers du dataset demandé sont déjà présents sur le site destinataire, ils ne sont pas transférés. Les transferts sont gérés en fonction

des politiques de transfert définies via les canaux FTS (File Transfer System) [17] qui respectent le modèle de calcul d'ATLAS. DQ2 possède des interfaces lui permettant d'opérer indifféremment sur les 3 types de grille (EGI, OSG, NorduGrid) utilisés par les sites d'ATLAS.

Le placement des données : La première année de prise de données, le modèle hiérarchique correspondant au modèle de calcul d'ATLAS a été appliqué. Une copie des données RAW est stockée au CERN sur bande et une autre est distribuée aux T1s en les gardant sur disque le plus longtemps possible. Pour les ESD, une copie est faite au CERN sur bande et 2 autres sont distribuées dans les T1s et gardées sur disque. Les AOD sont distribuées dans les T1s (2 copies) et les T2s (10 copies), la version précédente des AOD étant elle-même conservée. Des formats dérivés contenant une sélection des évènements des formats précédents sont aussi produits et distribués : les dESD contiennent les données utilisées par les groupes de performance, 10 copies de ces données sont distribuées aux T2s. De même les dAOD ainsi que les D3PD (données sous format ROOT [18]) destinés aux analyses de physique sont construits par les groupes de physiques et stockés dans des espaces réservés à ces groupes. Ainsi, la distribution des données au début du Run 1 se faisait selon le schéma de gauche de la figure 1 : les T2s reçoivent des données uniquement via le T1 du nuage auquel ils appartiennent.

Au vu de l'évolution de l'utilisation réelle des données, de leur accumulation et des progrès dans les fonctionnalités de la grille de calcul, la collaboration ATLAS a fait évoluer son modèle de calcul en continu pour en améliorer les performances. Certaines de ces évolutions ainsi que quelques exemples d'utilisation qui les ont déclenchées sont détaillés dans les points suivants :

– Les évènements au format RAW, en général stockés sur bande car volumineux, n'étaient donc pas directement accessibles. Afin de pouvoir étudier en détail et très rapidement les collisions en format RAW en cas de découverte, il a été décidé de garder les évènements dans ce format sur disque. Pour ce faire, les données RAW ont été compressées ce qui a permis de gagner un facteur 2 sur leur taille ; en même temps, le nombre de répliques et le temps de vie des ESD sur disque a été réduit.
– Pour suivre la progression de la luminosité instantanée du LHC et élargir les possibilités d'analyse tout en permettant un meilleur contrôle des systématiques, la collaboration ATLAS a décidé de doubler le taux d'acquisition des données de 200 à 400 Hz, ce qui rend nécessaire une meilleure gestion des espaces de stockage en particulier. En 2012, ce taux a même dépassé les 600 Hz avec la possibilité d'enregistrer 200 Hz de données supplémentaires dont le traitement est différé jusqu'au premier long arrêt du LHC (LS1).
– Les données pré-placées dans les sites ne correspondaient pas toujours aux données les plus demandées. Par exemple au début de la prise de données en 2010, contrairement au modèle, les ESD étaient intensivement utilisées plutôt que les dESD car il était nécessaire de bien maîtriser les données avant de pouvoir réduire leur format. Ainsi beaucoup de dESD stockées à l'avance sur les T2s n'étaient pas ou peu utilisées.

Pour répondre à ces problématiques, le système de distribution des données a été repensé et ne repose plus uniquement sur une distribution prédéfinie des données en fonction du modèle de calcul. Il utilise dorénavant aussi un placement des

données dynamique et sur demande. Ce placement (PD2P ou PanDA Dynamic Data Placement [19]) est lié au système d'analyse distribuée : lorsqu'un utilisateur demande l'accès à un dataset via une tâche d'analyse, ce dernier est automatiquement répliqué si nécessaire. Ainsi, les concepts de répliques primaires et secondaires ont été introduits. Les répliques primaires sont les données dont ATLAS garantit la présence sur disque selon le modèle de calcul. Les répliques secondaires sont des copies supplémentaires qui varient en fonction de la popularité des données et de l'espace disque disponible. Pour pouvoir gérer les données de cette façon, un système [20] mesurant la popularité des datasets (proportionnelle au nombre d'accès pour ce dataset) a été mis en place, en fonction duquel les datasets sont répliqués ou effacés. Enfin les réplications sur demande permettent de copier des données pour un besoin spécifique et après validation par la collaboration.

La dernière évolution du modèle de distribution des données est lié à la fiabilité du réseau internet tel qu'il a été expérimenté la première année de la prise de données. Il s'est avéré en effet que, non seulement les liens entre T1 et T2s d'un même nuage étaient de très bonne qualité, mais aussi que les liens entre de nombreux T2s et les T1s voir les T2s d'autres nuages l'étaient aussi. Il a donc été mis en place un système de test (dit SONAR) qui mesure les taux de transferts entre sites. Lorsque les taux de transferts entre un T2 et tous les T1s d'ATLAS sont supérieurs à un seuil, le T2 est labellisé T2D. Pour les sites T2D, les transferts via les T1s ne sont plus un passage obligé. Un algorithme de routage a été mis en place qui permet de choisir le chemin pour transférer les données en fonction des performances mesurées. Ce chemin doit passer par le T1 du nuage pour un T2 standard mais peut être plus direct pour un T2D en fonction du résultat de l'algorithme. Le système a donc évolué selon la figure 1 d'un schéma hiérarchique (à gauche) vers un maillage plus général (à droite).

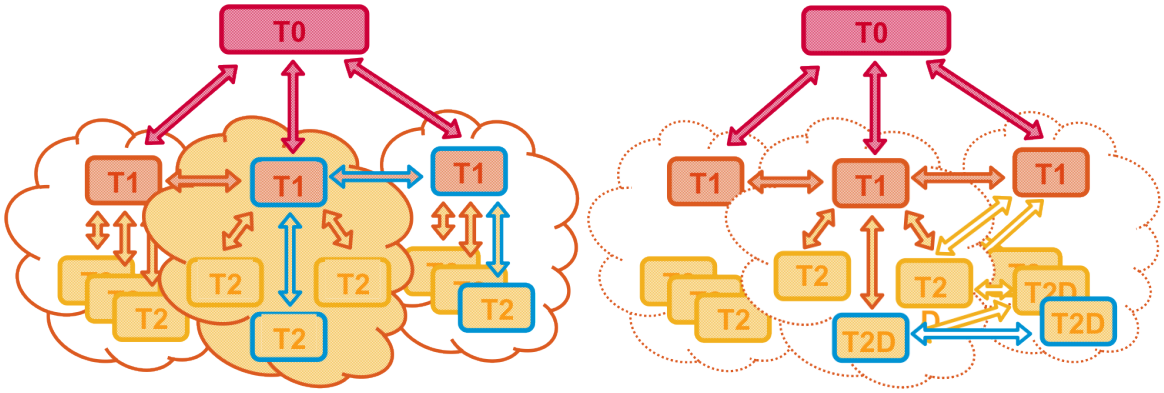


FIGURE 1 – Les transferts de données : schéma de principe du modèle hiérarchique (à gauche) et de son évolution (à droite).

– **La gestion des tâches de calcul**
Qu'il s'agisse de l'analyse ou de la production (génération, simulation, reconstruction des données), PanDA « Production and Distributed Analysis system » [21] est le système développé par ATLAS pour gérer et distribuer les différentes tâches de calcul sur la grille. On notera d'autre part, que pour les tâches d'analyse, un système alternatif Ganga [22] propose une interface qui permet leur gestion et leur envoi vers les grilles de calcul soit via PanDA soit directement vers les différents intergiciels des grilles uti-

lisées par l'expérience.

Les tâches de production et d'analyse sont soumises au serveur PanDA (voir figure 2) via une interface Python ou http et entrent en queue dans le système. Les tâches d'analyse sont envoyées au serveur directement par l'utilisateur, les tâches de production sont enregistrées dans une base de données (Production DB) et soumises au serveur PanDA automatiquement par l'intermédiaire d'un serveur nommé Bamboo. Le serveur PanDA distribue ensuite les tâches aux différents nuages en fonction de leur type, de leur priorité, des données d'entrées nécessaires, de l'emplacement et de la disponibilité de ces dernières, de l'espace libre disponible aux T1s ainsi que des ressources CPU utilisables (nombre de CPU réellement utilisés) et de la répartition prédéfinie entre les nuages (selon le protocole d'accord de WLCG).

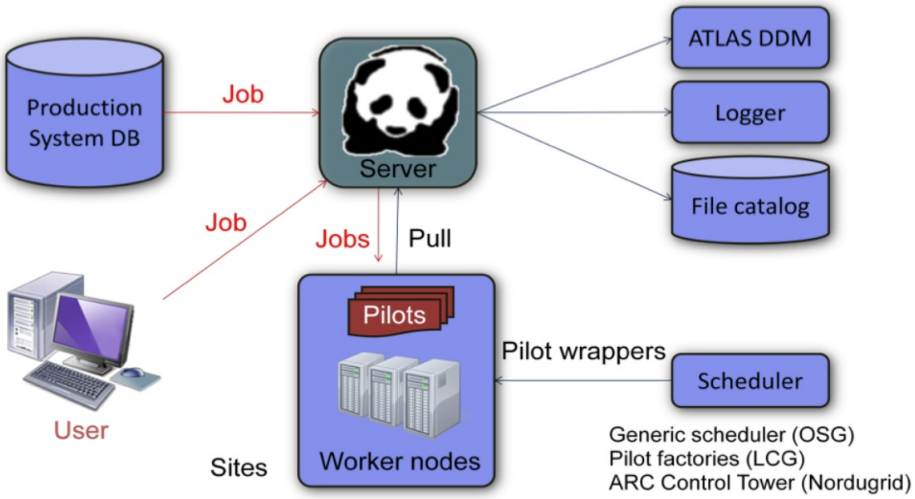


FIGURE 2 – Schéma de principe de soumission des tâches via PanDA.

Chaque tâche est divisée en un certain nombre de sous-tâches (jobs) qui permettent son exécution en parallèle. Une sous-tâche est l'unité de tâche informatique soumise aux nœuds de calcul des sites.

La répartition des sous-tâches dans le nuage se fait en fonction des caractéristiques de ces dernières : s'il s'agit de tâches nécessitant de nombreuses entrées/sorties ou la lecture de données sur bande, ou encore l'accès aux bases de données pour la reconstruction, elles seront attribuées au T1 du nuage. Dans le cas contraire, elles seront envoyées soit au T1 soit à l'un des T2 en fonction de la disponibilité de la version du logiciel d'ATLAS, de l'espace disque disponible dans le site, des caractéristiques techniques des nœuds du site (espace scratch, mémoire ...), de la disponibilité du site et de son taux d'occupation (nombre de sous-tâches en exécution/nombre de sous-tâches en queue).

Le traitement des tâches d'analyse est légèrement différent de celui des tâches de production. En effet, l'analyse étant en général gourmande en entrée/sortie, ses sous-tâches sont envoyées dans les sites qui hébergent les données dont elles ont besoin. Par contre pour la production, elles sont attribuées à un site qui ne possède pas nécessairement les données d'entrée. PanDA envoie alors une requête à DDM afin que les données nécessaires (datasets) aux sous-tâches soient disponibles sur le SE du site. DDM gèrera alors soit le transfert des données vers le site T2 soit, dans le cas où les

données sont stockées sur bande dans les T1, leur chargement sur disque. Pendant ce temps, les sous-tâches restent en queue dans le serveur PanDA (état assigné ou assigned) jusqu'à ce que DDM confirme au serveur que les données sont disponibles sur le site. Les sous-tâches passent alors à l'état activé (activated).

Les tâches ne sont pas directement soumises aux sites. Parallèlement au travail du serveur PanDA, un ordonnanceur (« scheduler »), envoie en permanence de petits programmes légers, les « pilot wrappers » [23] aux CE des sites. Pour les grilles OSG et EGI, l'ordonnanceur est appelé usine à pilotes (« pilot factory » [24]) et utilise Condor-G [25] comme système de gestion. Le CE du site via son système batch envoie le pilot wrapper sur un nœud de calcul (noté WN pour Worker Node) où il charge le code du pilote de PanDA. Le pilote vérifie dans un premier temps l'environnement du WN, ainsi que son espace et sa mémoire réellement disponibles et demande au serveur PanDA une sous-tâche en adéquation avec les caractéristiques du WN et du site. S'il n'y a pas de sous-tâche disponible pour la configuration du WN, le pilote s'arrête. Si une sous-tâche est attribuée, le pilote met en place l'environnement nécessaire (logiciels, librairies, bases de données) puis copie depuis le SE du site les données d'entrée sur le WN. Le code de la tâche de production ou de l'utilisateur est alors exécuté et le pilote suit son bon déroulement et en rend compte au serveur PanDA. Pendant cette phase, ce dernier assigne à la sous-tâche l'état « running ». Une fois celle-ci terminée, les fichiers de sortie sont enregistrés sur le SE du site et le pilote informe le serveur PanDA de la fin du programme. Enfin le pilote nettoie le WN et s'arrête. Si les données de sortie doivent être transférées (par exemple pour les tâches de production du T2 vers le T1 ou pour celles d'analyse d'un T2 vers un autre), DDM prend en charge ce transfert. L'état de la sous-tâche est alors « transferring ».

L'utilisation de pilotes permet ainsi d'utiliser efficacement le CPU disponible puisque le programme ne commence à tourner que lorsque le WN a la capacité d'accueillir la sous-tâche et lorsque les données sont disponibles et l'environnement en place et il s'arrête dès la fin du traitement. Les éventuels transferts de données de sortie se font dans un deuxième temps. Les pilotes permettent aussi de cacher à PanDA l'hétérogénéité des différentes grilles et des sites. Pour ce faire, le pilote obtient les configurations des sites via une base de données appelée PanDA SchedConfig DB. Les informations contenues dans cette base de données sont par exemple l'emplacement des logiciels d'ATLAS dans les sites, les chemins pour mettre en place l'environnement du programme, les outils utilisés pour copier les fichiers (les pilotes utilisent 21 outils de copie différents !), la possibilité d'accéder directement au système de stockage (par exemple via xrootd [26]), les caractéristiques des WN (espace et mémoire) ... On notera que l'ajustement du nombre de pilotes envoyés au site est parfois délicat, il doit prendre en compte la taille du site, être suffisant pour permettre d'écouler le plus rapidement possible les tâches en attente dans PanDA et ne pas être trop nombreux pour ne pas asphyxier le site et utiliser inutilement des créneaux qui pourraient être utilisés par d'autres VO.

– **Les bases de données**
Les bases de données sont omniprésentes et essentielles dans le traitement des données d'ATLAS et se retrouvent à tous les niveaux depuis le traitement en ligne des données jusqu'à leur analyse. Je n'entrerai pas ici dans le détail de leur fonctionnement. On notera simplement que la plupart des bases de données utilisées par l'expérience fonc-

tionnent sous Oracle.

La base de données « Conditions Database » [27] regroupe les conditions dans lesquelles ont été prises les données et leur validité dans le temps (informations sur le faisceau, l'état des détecteurs, la luminosité, le système de déclenchement ...). Pour traiter les données, il est nécessaire de connaître la géométrie des détecteurs, leurs données d'étalonnage et la qualité des données. Certaines de ces informations ne sont renseignées qu'après l'analyse d'une partie des données. Il peut exister plusieurs versions de ces informations. À chaque nouveau traitement des données, les versions des logiciels et des étalonnages utilisés etc., sont référencées de sorte que la traçabilité des données analysées soit totale. Ainsi, pour chaque fichier ou entité lié aux données d'ATLAS, les bases de données contiennent des méta-données (des données sur les données), permettant de retrouver les conditions dans lesquelles ces données ont été prises puis traitées. On notera qu'une interface web AMI [28] a été conçue pour accéder facilement aux différentes méta-données liées à un fichier ou à un dataset.

Bien entendu, d'autres bases de données sont nécessaires aux traitements des données et au fonctionnement de la grille :
- la base de données liée aux versions des logiciels d'ATLAS,
- les catalogues centraux qui concernent les fichiers (ils contiennent la liste des fichiers avec leur GUID, LFN, taille, checksum, ainsi que leur lien avec les datasets qui les contiennent), les datasets (avec les noms des datasets et leurs attributs : possesseur, date de création, état, versions ...), les répliques des datasets,
- la base de données de PanDA qui permet de suivre le tâches de traitement des données,
- les bases de données LCG liées aux sites de la grille de calcul,
- ...

Quasiment chaque entité de la grille est liée à une base de données.

- **Les outils de monitoring et de surveillance**

  Il serait difficile de faire fonctionner un système aussi complexe que la grille de calcul d'ATLAS et encore moins de l'améliorer sans des outils de monitoring et de surveillance performants. Le monitoring a été ainsi développé et amélioré au fil du fonctionnement de l'expérience et couvre l'ensemble des aspects de la grille ATLAS (voir la figure 3) : les activités du T0, le traitement des données via PanDA pour la production et l'analyse, les transferts de données, le remplissage des disques et le nombre de CPU utilisés par site, l'état des sites et des bases de données ... L'historique de ces données est aussi conservé, ce qui permet une analyse a posteriori du fonctionnement de tous les aspects de la grille de calcul. Ceci est très précieux pour étudier les améliorations utiles. Ces outils sont disponibles via des interfaces web facilement accessibles. L'ensemble des outils de monitoring sont décrits plus en détail dans la référence [29].

- **Les équipes**

  Dans chaque secteur de la grille de calcul, des équipes travaillent sur le développement et l'amélioration des outils nécessaires à son fonctionnement et d'autres s'assurent que le fonctionnement au jour le jour de cet ensemble complexe soit le meilleur possible. Il est difficile de faire le compte du nombre de personnes travaillant sur la grille de calcul pour ATLAS. Il faut compter les personnes qui gèrent les sites, celles qui

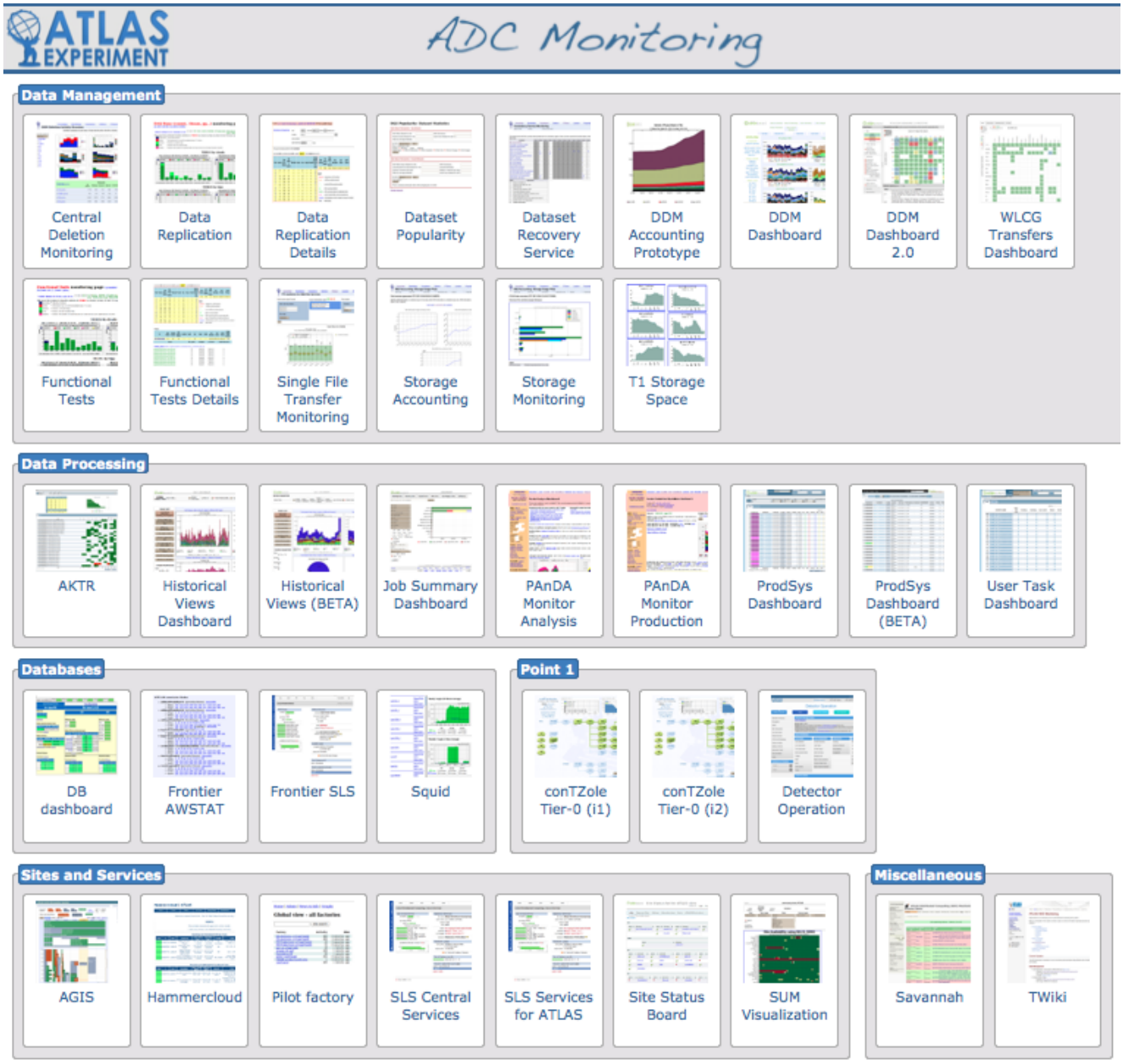


FIGURE 3 – Capture du site web regroupant les principales pages du monitoring de l'expérience ATLAS (http ://adc-monitoring.cern.ch/).

travaillent sur les outils propres à ATLAS ou communs aux expériences du CERN en passant par celles qui développent les intergiciels au sein de LCG. En 2011, le nombre de périodes de support (shift) pour l'informatique distribuée représente environ 10 % du support nécessaire pour faire fonctionner l'ensemble de l'expérience. Par ailleurs, les personnes travaillant pour l'ensemble du traitement informatique pour ATLAS représentent environ 150 équivalents temps plein dans la collaboration (environ la moitié participent aux activités liées au calcul distribué), auxquels il faut ajouter environ 200 équivalents temps plein dans les sites.

Je liste ici les équipes qui assurent le bon fonctionnement de la grille :

- l'équipe Comp@P1, localisée au CERN, assure le suivi du traitement des données au T0 ainsi que le transfert des données vers les T1s et les quelques T2s utilisés pour l'étalonnage des détecteurs.
- l'équipe ADCoS (ATLAS Distributed Computing Operations Shift) assure un support 24h sur 24 et 7 jours sur 7, pour le traitement et la distribution des données pour la centaine de sites qui contribuent à ATLAS.
- une équipe Squad par nuage fait le relais entre les sites appartenant à leur nuage et la collaboration ATLAS et ses différentes équipes ; ses membres (dont je fais partie) assurent le suivi des sites au jour le jour et vérifient que toutes les tâches en cours sont correctement traitées. Tous les aspects de ce traitement sont pris en compte : stockage, transfert de données, accès aux logiciels d'ATLAS, traitement des données en lui-même ... Le Squad est aussi responsable des services hébergés dans les nuages. Il s'assure par exemple du bon fonctionnement de l'usine à pilote associée au nuage. Il gère aussi la déclaration et la définition des sites et de leurs paramètres pour les outils d'ATLAS. L'équipe du Squad intervient en cas de problème lié à un site de son nuage ou à un outil d'ATLAS défectueux vis à vis de celui-ci. Elle permet de rendre compte rapidement des éventuels problèmes qui auraient un impact pour la production centralisée ou les tâches d'analyse d'un utilisateur et aide les administrateurs de site à résoudre les problèmes liés à l'expérience. L'équipe s'assure que les sites de son nuage reçoivent bien les données et les tâches qui doivent leur être attribuées. D'autre part, elle suit les performances des sites sur le long terme et travaille avec leurs administrateurs pour optimiser au mieux leur utilisation. Elle relaie les demandes d'évolution d'ATLAS vers les sites et assure le retour d'expérience de ceux-ci vers la collaboration ATLAS.
- l'équipe DAST (Distributed Analysis Support Team) se relaie pour assurer un support aux utilisateurs qui analysent les données en fonction de la zone géographique à laquelle ses membres appartiennent (Asie-Pacifique de 0 à 8 h HNEC, Europe de 8 à 16 h HNEC, Amérique de 16 à 24 h HNEC).
- enfin un expert de l'informatique distribuée est de permanence pour faire remonter les problèmes particuliers aux services centraux d'ATLAS lorsque nécessaire.

1.3.4 Les évolutions récentes

Comme je l'ai déjà mentionné au fil des paragraphes précédents, l'ensemble de l'infrastructure et des logiciels de traitement des données d'ATLAS n'a cessé d'évoluer [30], d'une part parce que la technologie informatique évolue elle aussi en continu, et d'autre part parce que l'accumulation des données et l'expérience engrangée depuis le démarrage du LHC entrainent de nouvelles demandes et déclenchent de nouvelles améliorations. On notera que

ces évolutions ont été accompagnées par le développement et l'amélioration des outils de monitoring nécessaires à la compréhension du fonctionnement du système et à la mise en évidence et au diagnostic des problèmes et de leur fréquence. Ceci permet alors la mise en place de solutions efficaces ainsi que leur évaluation.

L'évolution la plus importante a déjà été mentionnée et est liée à la qualité du réseau qui a permis de faire tomber les frontières entre nuages. Ainsi les données peuvent transiter directement d'un T2 à un autre sans passer par deux T1s comme il était prévu à l'origine. Cela permet de profiter au mieux des infrastructures réseau existantes, d'économiser du disque dans les T1s et de diminuer les sources d'erreurs en éliminant les intermédiaires. De plus, les T2s qui possèdent une bonne connexion aux autres T1s peuvent dorénavant faire de la production pour les autres nuages ; c'est-à-dire que les données produites par un T2 sont transmises à un autre T1 que celui de son nuage ; cela permet ainsi de mieux répartir la charge des différents sites, tout en respectant la répartition des données telle qu'elle a été anticipée.

Bien sûr, ces évolutions induisent une utilisation plus intensive du réseau et les expériences LHC ont réfléchi à son évolution avec les fournisseurs du réseau (RENATER [31] en France). Ainsi la création d'un réseau spécifique aux sites utilisés par le LHC, LHCONE [32], a été décidée et est en cours de développement. Ce réseau permettra d'améliorer la qualité et le taux des transferts des données LHC entre sites en isolant ces transferts du reste des mouvements de données sur internet. Cela permettra de limiter l'impact des grands flux de données du LHC sur le réseau général et à l'inverse de protéger les transferts LHC des variations des transferts généraux et éventuellement d'attaques malveillantes.

Le deuxième axe de travail a été d'améliorer la fiabilité du système en éliminant les problèmes ou à défaut en les détectant au plus vite pour éviter de soumettre de nouvelles tâches à un système qui ne fonctionne pas. En effet, même si le taux de fiabilité de chaque site ou système est très grand, le nombre de ces sites ou systèmes nécessaires à chaque tâche étant lui aussi très grand, le taux global d'erreur n'est pas négligeable. Ainsi, 98 % du traitement d'une tâche est très rapide mais la gestion des quelques pour-cent restant peut vite s'avérer longue et fastidieuse. Ceci impacte particulièrement les tâches d'analyse, les tâches de production étant mieux automatisées et moins diversifiées donc plus facilement adaptables aux ressources.

Un exemple de problème résolu est la distribution dans les sites des différentes versions du logiciel et des bases de données nécessaires à la reconstruction des données. Dans chaque site, les différentes versions du logiciel étaient installées et testées par un système dédié et les extraits des bases de données nécessaires aux T2s, enregistrés dans des fichiers distribués via DDM. Il a alors été observé que de nombreuses tâches pouvaient accéder simultanément aux bases de données ce qui représentait un goulot d'étranglement. D'autre part, la probabilité qu'un site au moins n'ait pas une version du logiciel correctement installée était relativement importante. Pour palier ces problèmes, ATLAS installe dorénavant le logiciel et les bases de données uniquement au CERN. Le logiciel est obtenu par les sites via CERNVM-FS [33] et les bases de données via les systèmes Frontier/Squid [34] qui permettent de copier les données dans un cache par http. Cette évolution a grandement amélioré la fiabilité de l'accès au logiciel d'ATLAS et aux bases de données et permet même à tous les T2s d'avoir accès aux bases de données jusque-là dupliquées uniquement dans les T1s. Ils peuvent ainsi reconstruire les données réelles ce qui n'était possible que dans les T1s auparavant.

Tous les problèmes ne peuvent malheureusement pas avoir de solution globale immédiate. Dans ce cas, la collaboration a travaillé pour surveiller et interdire l'accès aux ressources qui ne fonctionnent pas, puis pour les rendre accessible à nouveau dès que le problème a été résolu. Il s'agit ainsi de surveiller le système pour détecter la survenue d'erreur à un taux anormal, d'analyser l'erreur, de sortir de production le système fautif et de prévenir les personnes qualifiées pour corriger le problème. Une fois le problème résolu, le système est testé avant d'être remis en production. Dans un premier temps, les actions nécessaires à ce traitement ont été menées manuellement par les équipes de l'informatique distribuée et elles sont maintenant peu à peu automatisées. Actuellement par exemple, des tâches de tests, étalonnées et validées, et donc qui doivent fonctionner, sont envoyées régulièrement sur tous les sites. Si plus qu'un certain nombre de ces tâches successives sont en erreur, le site est automatiquement sorti de production. Les tâches de test continuent alors de sonder le site et dès que les traitements des tâches redeviennent positifs, le site est remis en ligne. L'intervention humaine reste tout de même nécessaire pour comprendre les problèmes et trouver des solutions sur le long terme. Cette automatisation a permis d'améliorer de façon importante la fiabilité du système.

1.3.5 Performances et résultats

En 2010, 2011 et 2012, l'expérience ATLAS a enregistré respectivement 45 pb^{-1} et 5,2 fb^{-1} avec des collisions à 7 TeV dans le centre de masse puis 21,7 fb^{-1} à 8 TeV soit un total correspondant à 26,9 fb^{-1}. L'ensemble représente quelques 5 milliards de collisions réelles auxquelles il faut ajouter les collisions simulées.

En 2012 par exemple, environ 4 milliards de collisions ont été simulées (un peu moins de la moitié en simulation rapide) à un rythme d'environ 50 millions par jour. Environ 800 utilisateurs utilisent la grille de calcul qui pour ATLAS correspond à 130 sites. L'ensemble a fonctionné avec un très bon niveau de performance et a su s'adapter à des demandes et des conditions qui n'avaient pas toutes pu être anticipées : grands taux de déclenchement (le double de ce qui avait été initialement prévu), d'évènements d'empilement, une production Monte Carlo intense et une forte demande des utilisateurs pour l'analyse.

Quelques chiffres et graphes sont présentés ci-dessous pour donner une idée du niveau d'activité et de performance de la grille pour ATLAS ainsi que de la progression de son utilisation depuis le début du Run 1 :

- depuis le démarrage, le nombre de tâches simultanées en exécution n'a cessé d'augmenter pour dépasser 140 000 tâches en 2012 (voir la figure 4) ; un peu plus de 50 % d'entre elles sont des tâches d'analyse ;
- la consommation en CPU en 2012 est de 3 000 milliards d'heures HEP-SPEC06 (ou 4 millions d'années de calcul sur une machine actuelle !) ; l'analyse ne représente que 22 % du CPU total alors que la simulation, le traitement des évènements d'empilement, la génération d'évènements Monte Carlo et la reconstruction représentent respectivement, 41, 17, 9 et 6 % du total ;
- en 2012, l'efficacité moyenne de succès pour l'ensemble des tâches est de 90 %. Pour les tâches d'analyse, cette efficacité est de plus de 80 %, pour les tâches de production elle varie en moyenne de 80 % pour la génération d'évènements à 95 % pour les étapes de simulation du détecteur. D'autre part, les efficacités CPU sont aussi remarquables : 80 % pour l'analyse et plus de 90 % pour la production ;
- fin 2012, les données toutes catégories confondues (données réelles, simulées, données

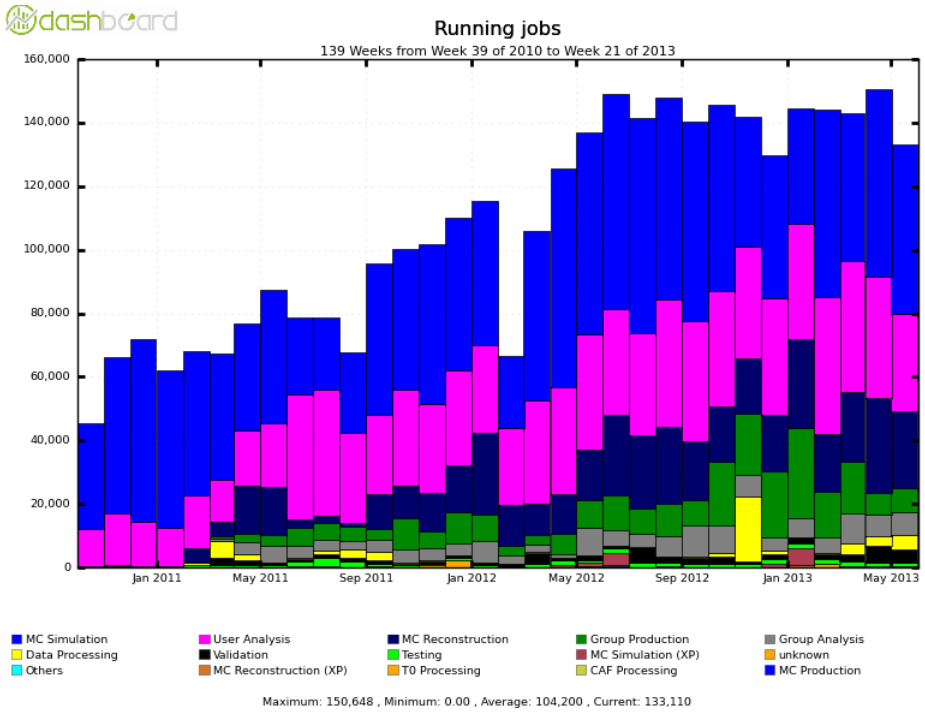


FIGURE 4 – Nombre de tâches simultanées exécutées sur l'ensemble des sites de la grille pour l'expérience ATLAS depuis octobre 2010.

après traitement pour analyse), sur disque et sur bande représentent environ 130 Po (voir figure 5) ;
- ces données correspondent à environ 400 millions de fichiers ;
- le taux d'accès aux fichiers est de l'ordre de 5 millions par jour ;
- le taux de transfert du T0 vers les T1s est autour de 2 Go/s en moyenne sur une journée et sur l'ensemble de la grille d'ATLAS, le taux de transfert global est de 10 Go/s ; environ 20 Po de données transitaient par mois sur la grille fin 2012. La figure 6 donne l'évolution des transferts depuis 2009.

Utilisation et performances du nuage français

Un peu plus de 12 % des tâches d'ATLAS ont été traitées sur le nuage français soit 37 millions de tâches sur l'année et plus de 15 000 tâches tournent en continu sur le nuage. La répartition du nombre de tâches réalisées en 2012 par nuage et par site pour le nuage français sont montrés sur la figure 7.

L'efficacité moyenne du nuage est similaire à celle de l'ensemble des nuages. La figure 8 donne l'efficacité c'est à dire le pourcentage de tâches terminées avec succès pour chaque site. Ce nombre est donné à titre indicatif, car tous les sites ne reçoivent pas tous les mêmes types de tâches, qui n'imposent pas les mêmes contraintes et possèdent en quelque sorte une efficacité intrinsèque différente. Pour comparer l'efficacité des sites, il est plus juste de le faire par type d'activité, c'est ce qui est montré sur la figure 8 pour les tâches de reconstruction. Les inefficacités sont dues à différents problèmes. Les 3 types d'erreurs les plus fréquentes pour la reconstruction sont les problèmes d'accès aux données, les arrêts des tâches par le gestionnaire de tâches des sites (durée de la tâche trop longue, trop grande consommation de mémoire, autres problèmes) et les problèmes d'accès au logiciel d'ATLAS.

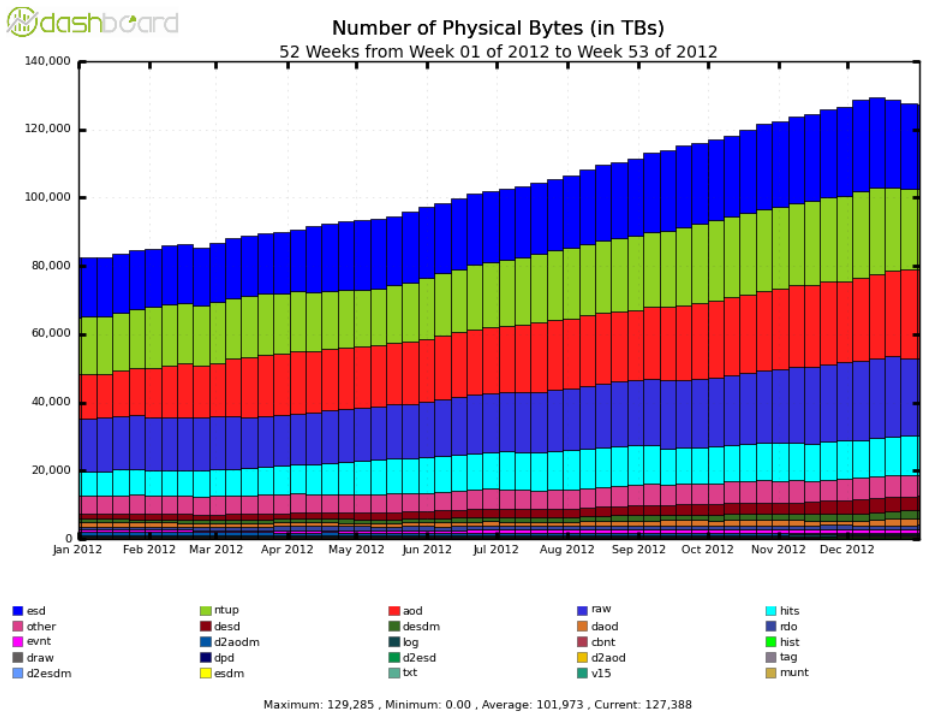


FIGURE 5 – Quantité de données enregistrées sur la grille pour l'expérience ATLAS en fonction du temps pour l'année 2012.

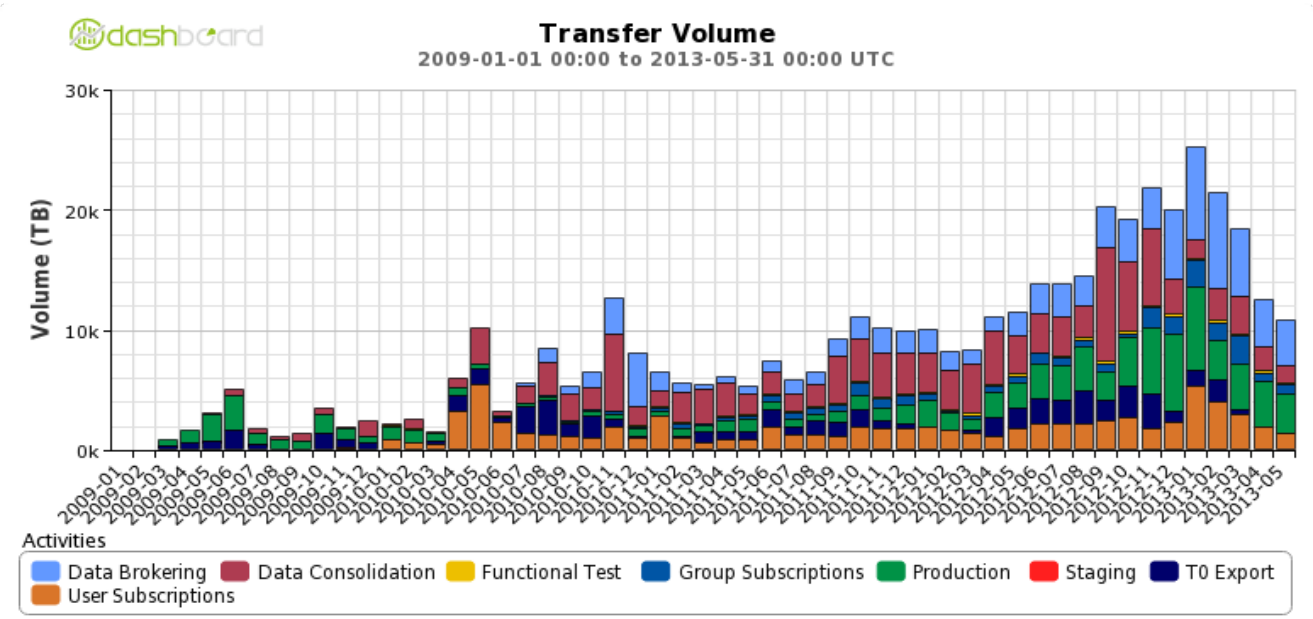
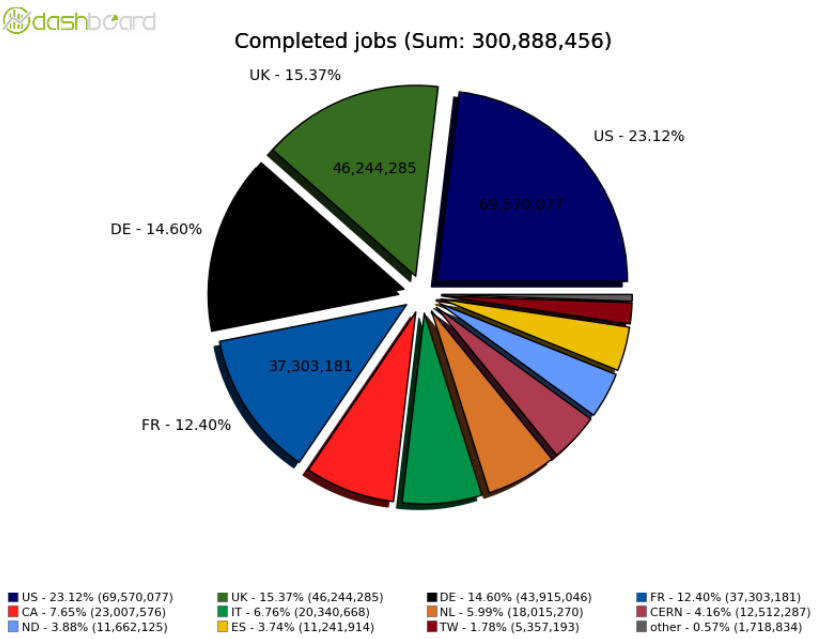


FIGURE 6 – Évolution de la quantité de données transférées pour l'ensemble de la grille pour l'expérience ATLAS depuis 2009.



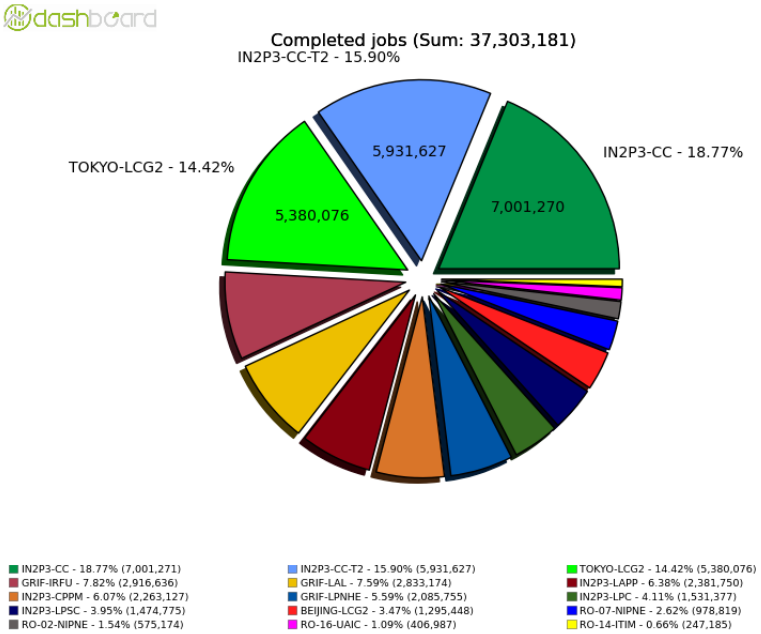


FIGURE 7 – Répartition du nombre de tâches effectuées pour l'année 2012 en fonction des nuages (en haut) et des sites pour le nuage français (en bas).

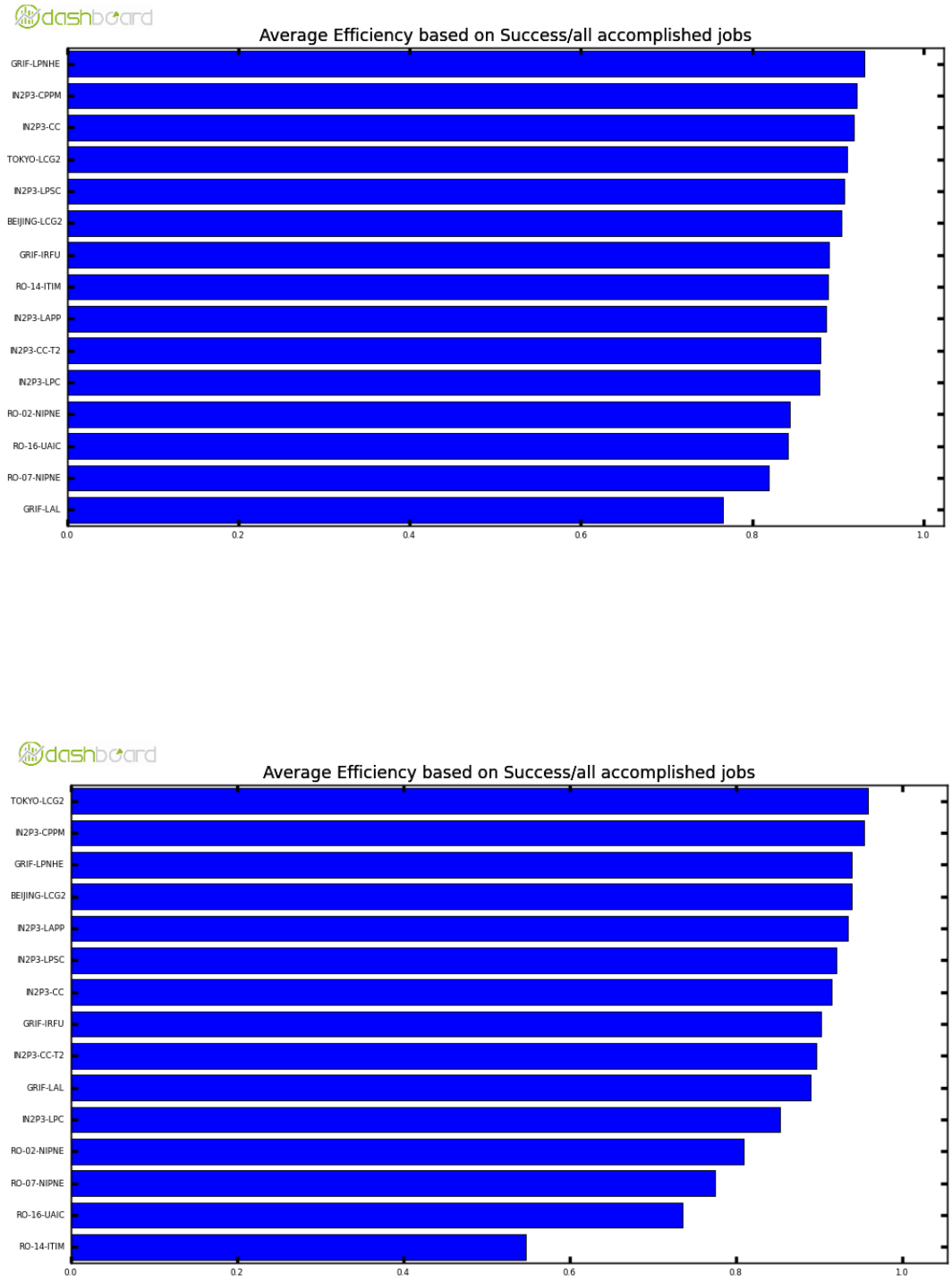


FIGURE 8 – Efficacité des sites du nuage français en 2012 (pourcentage de tâches terminées avec succès) pour l'ensemble des types de tâches en haut et pour les tâches de reconstruction en bas.

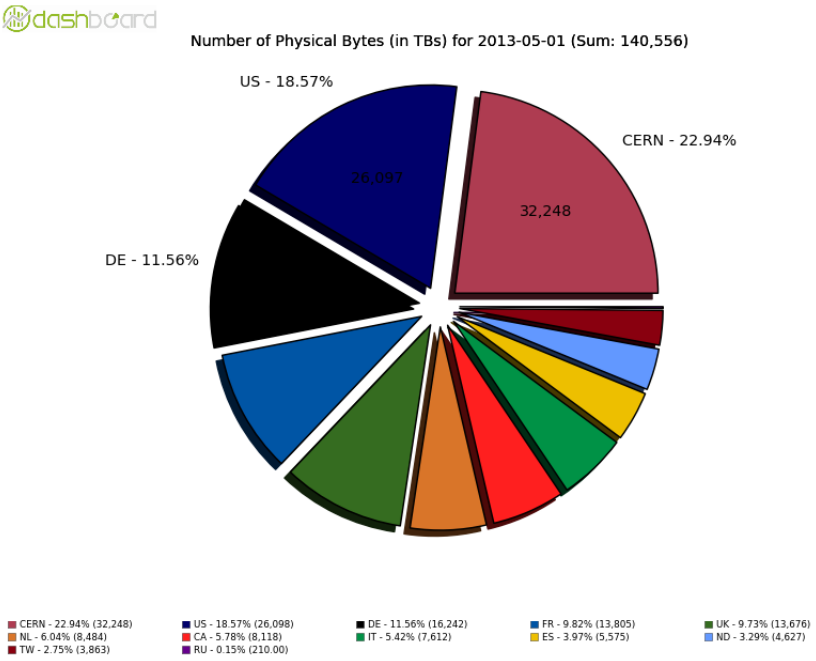
En ce qui concerne les données, le nuage français accueille environ 10 % des données de l'expérience, leur répartition par nuage et par site pour le nuage français est donnée sur la figure 9.

1.3.6 Perspectives

La collaboration ATLAS continue à développer ses outils informatiques. Le premier long arrêt du LHC (LS1) entre 2013 et 2015 est l'occasion de travailler sur des améliorations majeures de ces derniers. En effet, ceux-ci ont été conçus en général quelques années avant le démarrage du LHC et fonctionnent avec succès depuis 7 ou 8 ans. Le long arrêt du LHC est donc une période propice pour développer de nouveaux outils pour les années qui viennent permettant d'aller au delà des limitations de leurs prédécesseurs, de répondre aux demandes d'évolutions et de nouvelles fonctionnalités de l'expérience tout en exploitant le développement des technologies récentes. Ainsi, de nouveaux logiciels pour la gestion des tâches et des données sont en cours de développement et de tests et devraient être mis en production prochainement. Cependant, la deuxième phase de prise de données du LHC présente plusieurs défis à la communauté informatique. L'objectif de l'expérience ATLAS en effet est de prendre des données avec un taux de déclenchement doublé par rapport à la phase précédente soit 1 kHz et avec des conditions de prise de données au moins aussi complexes que précédemment (empilement des évènements, taille et complexité des évènements, ...). Les ressources techniques et humaines sont contraintes par un budget qui devrait rester stable, ce qui devrait permettre un accroissement des ressources de l'ordre de 20 à 30 % pendant l'arrêt du LHC. La seule façon de relever le défi est donc d'améliorer significativement le logiciel de traitement des données et l'organisation des différentes phases de traitement pour éliminer les redondances existantes. La collaboration a commencé ce travail qui doit se poursuivre pendant la phase d'arrêt du LHC.

Le deuxième défi, qui contribuera certainement aussi à relever le premier, est de continuer d'adapter le traitement informatique des données à l'évolution des technologies et d'accompagner cette évolution. En effet, les progrès sont constants en ce qui concerne le réseau (avec la possibilité de construire des fédérations de stockage) et les processeurs (calcul parallèle, GPU ...). On note aussi l'apparition de nouveaux outils et services : la virtualisation, les clouds commerciaux ... l'ensemble de ces aspects sont pris en considération par l'expérience et certains sont déjà testés et mis en œuvre.

Enfin, la partie opérationnelle de la grille nécessite un temps de travail important de la part de nombreuses personnes. Pour minimiser ce temps, un effort continu a été fait qui concerne le développement du monitoring de l'ensemble du système ainsi que de son automatisation. Par ailleurs la centralisation de certains services comme la distribution des logiciels et des bases de données déployés au CERN et accédés via CERNVM-FS permet aussi de mutualiser les efforts. Ce travail devra se poursuivre et les nouveaux outils en développement prennent en compte les besoins de simplification de cette partie opérationnelle.



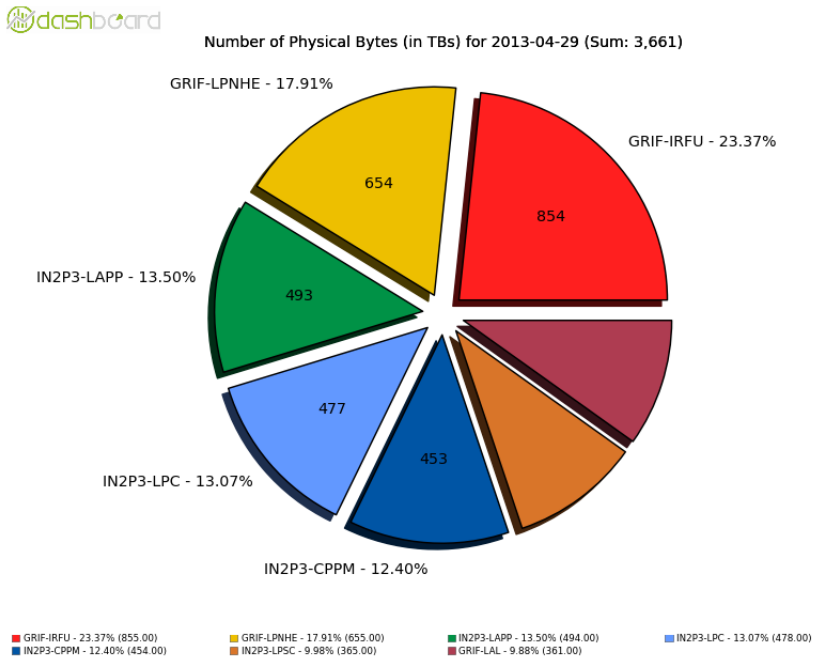


FIGURE 9 – Répartition des données début 2013 en fonction des nuages (en haut) et des sites pour le nuage français (en bas) sans le T1 (CC-IN2P3) qui représente environ 70 % du stockage français.

2 Le nœud de grille du LPSC

Le site du LPSC (IN2P3-LPSC) est l'un des 14 sites de grille français qui participent au calcul et au stockage LHC. En 2012, sans compter les ressources du centre de calcul de l'IN2P3 (T1 et T2), il possède 7 % des ressources de calcul et 9 % du stockage des T2s français. Il supporte parmi les VO LHC les expériences ATLAS et ALICE.

2.1 Bref historique

Après une petite participation à « datagrid » [35] en 2001, c'est en 2006 que le LPSC a décidé de créer un nœud de grille LCG suite à la demande des physiciens du groupe ATLAS. Le laboratoire investissait à l'époque dans l'aménagement d'une salle informatique avec un système de refroidissement original basé sur la technique du « free cooling [5] » et dans toute l'infrastructure nécessaire à l'accueil d'un nœud de grille. Après une phase de tests en 2007 et l'achat des premiers matériels soutenus par le LPSC, les groupes de physique LHC du laboratoire et l'Institut des Grilles, le site est mis en production en janvier 2008 en tant que T3 de la grille WLCG. Peu à peu les ressources du site s'accroissent et il s'ouvre à des VO hors LHC en 2009 et 2010.

L'expérience ALICE ne différencie pas les sites T3s des T2s dans leur utilisation. En ce qui concerne l'expérience ATLAS, le site a aussi fonctionné dès le départ comme un T2 grâce à son bon fonctionnement et à l'implication de son personnel. Pendant cette période, la principale différence avec les T2s est qu'aucune ressource n'était officiellement engagée pour les expériences. Cependant, au début de l'année 2010, avec l'arrivée de nombreux sites T3s avec des services de grille minimaux, la collaboration ATLAS a clairement différencié les activités des T3s et des T2s en particulier en matière de distribution des données mais aussi de tâches confiées. Les données n'étant plus automatiquement importées, l'activité d'analyse s'est donc restreinte aux utilisateurs locaux. Ceux-ci importaient les données des groupes de physique auxquels ils appartenaient ou leurs propres données d'analyse sur le site. Ainsi seuls les utilisateurs locaux ou extérieurs intéressés par ces données particulières utilisaient le site pour faire tourner leur programme d'analyse.

Cette évolution de l'utilisation des T3s par la collaboration ATLAS, la bonne qualité du site, l'implication de son personnel et la volonté du laboratoire et des groupes LHC de participer activement au calcul LHC sont les principales raisons qui ont mené à la demande de passage à un site de type T2. Ce passage, avec le soutien des expériences ATLAS et ALICE, a été accepté par WLCG à l'été 2011. Depuis le site a poursuivi sa croissance en terme de ressource et continue à accompagner les demandes des VO LHC, ALICE et ATLAS, tout en restant ouvert à d'autres disciplines.

2.2 Description du nœud de grille

Les serveurs du T2 du LPSC se trouvent dans la salle informatique du laboratoire qui permet l'hébergement de 8 baies 42 U avec une puissance électrique disponible de 120 kW.

5. La technique du free-cooling est simplement basée sur la circulation d'air extérieur pour refroidir les serveurs informatiques. Elle remplace avantageusement la climatisation puisqu'elle présente un coût de fonctionnement très faible et une fiabilité largement supérieure aux systèmes de climatisation habituels.

La salle dispose d'un onduleur (30 kVA) pour les services critiques et d'un système original de refroidissement par « free cooling » appelé Ecoclim [36] particulièrement fiable, d'une puissance de 100 kW. Le T2 occupe aujourd'hui 5 des 8 baies disponibles dans la salle informatique du LPSC. Sa capacité de calcul est constituée d'une centaine de serveurs, correspondant à 740 cœurs d'une puissance totale de 6 700 HEP-SPEC 06 . Quant au stockage, 27 serveurs offrent une capacité de 1 Po brut qui correspond à 700 TO net (utilisable via le « Disk Pool Manager » DPM). Les évolutions de la puissance de calcul et de stockage depuis 2008 sont représentées sur la figure 10.

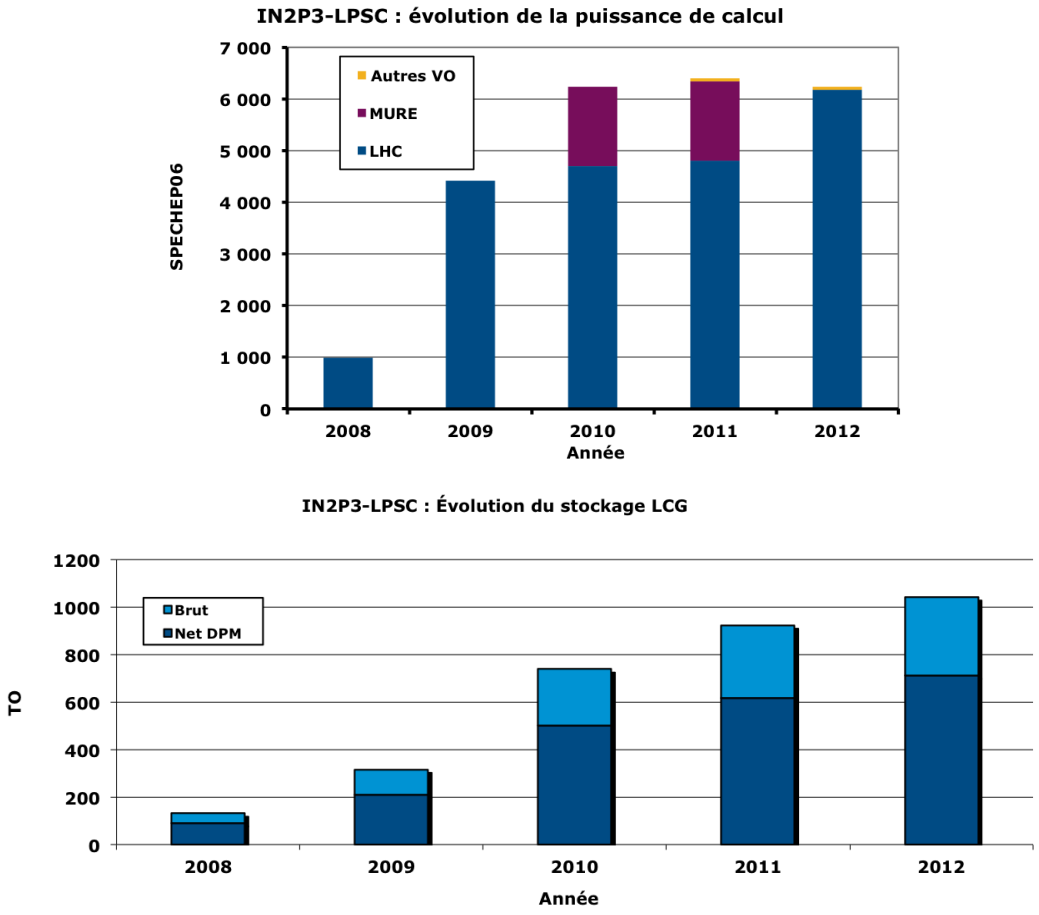


FIGURE 10 – Évolution de la capacité de calcul (en haut) et de stockage (en bas) du noeud de grille du LPSC (* en bleu foncé la capacité net utile telle que vue par DPM).

Les services de grille disponibles au LPSC sont répartis sur 6 serveurs avec une alimentation redondante. Le T2 du LPSC dispose ainsi d'un BDII « Database Information Index » pour le site, de 2 CE « Computing Elements » (CREAM-CE), de plusieurs « User Interface », d'une VOBOX, de 2 DPM « Disk Pool Manager » (xrootd ou non). Les configurations de l'ensemble des services sont déployées via Quattor [37]. L'ensemble de la grille est surveillé via le serveur NAGIOS [38] du LPSC et le système de sondes d'EGI.

2.3 Le réseau

Le cœur du réseau local est basé sur 2 switchs CISCO NEXUS 5010 dotés de 40 ports à 10 Gbits/s. Le LPSC étant situé sur le polygone scientifique de Grenoble, la connexion à RENATER dont le point de présence est sur le campus de Saint-Martin-d'Hères à une distance d'environ 10 km se fait via le réseau métropolitain Metronet. Le point d'accès de ce dernier pour le CNRS est géré par le laboratoire Louis Néel. Le débit maximum est passé de 2 Gbits/s à 10 Gbits/s fin septembre 2012, il est restreint à 5 Gbits/s pour les activités grille du laboratoire. Le point de présence RENATER est ensuite relié par fibre optique à 10 Gbits/s à Lyon, Genève et Cadarache. On peut voir la nette amélioration des taux de transfert vers les T1s d'ATLAS suite à l'amélioration du réseau en septembre sur la figure 11.

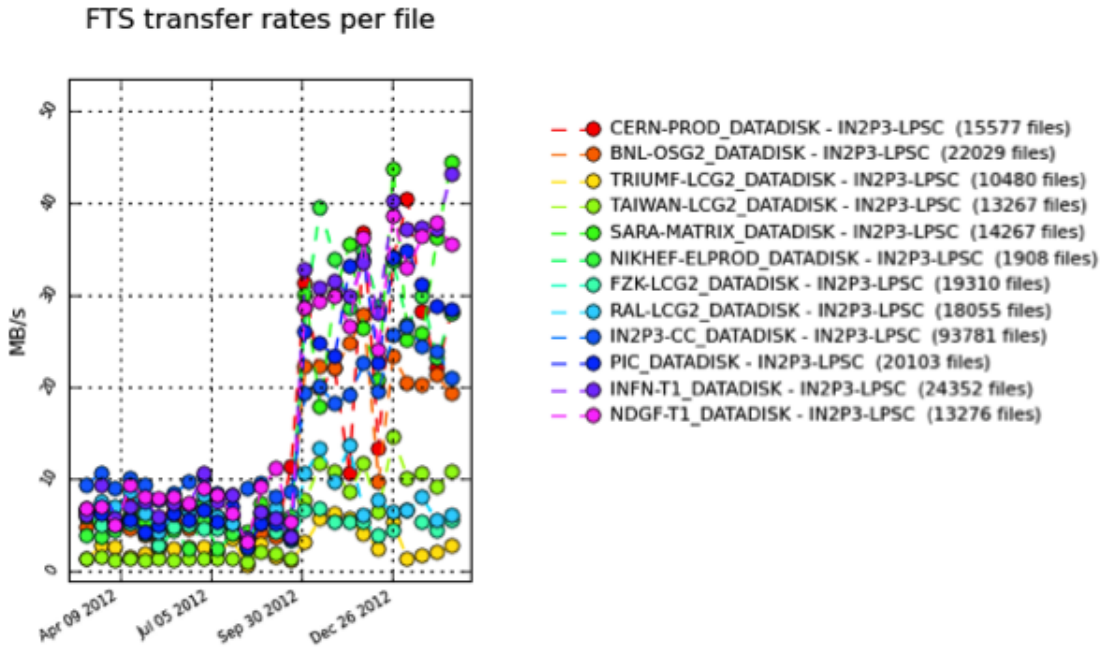


FIGURE 11 – Évolution des taux de transfert des T1s supportant la VO ATLAS vers le site du LPSC pour les fichiers de plus de 1 Go.

Cette mise à niveau du réseau permet d'améliorer les débits vers les sites hors du nuage français et de maintenir le statut de T2D du site. La qualité de T2D est réservée aux sites possédant une bonne connexion vers les 12 T1s [6] ainsi qu'une bonne qualité de service. Les sites T2D peuvent recevoir les données de tous les T1 directement en particulier pour le placement dynamique des données et faire de la production pour tous les T1s.

Le site du LPSC n'est pas encore sur le réseau privé LHCONE mais sa connexion est en cours d'étude. Elle permettra de garantir la qualité et la sécurité de la connexion vers les autres sites participant au calcul LHC. Ceci explique le fait que la connexion au site FZK par exemple n'a pas bénéficié de l'amélioration du réseau, FZK utilisant LHCONE comme gage de sécurité, filtre les données des sites hors LHCONE via un réseau secondaire.

6. le taux de transfert moyen sur une semaine des fichiers de plus de 1 GB doit être supérieur à 5 Mo/s sur 3 des 4 dernières semaines et pour au moins 9 des 12 T1s et T0 et ce dans les 2 sens.

2.4 Le personnel

Les activités du LPSC étant structurées sous forme de projet, la plateforme technique du T2 du LPSC possède un responsable scientifique (moi-même) et un responsable technique. Parmi les membres du service informatique qui permettent le bon fonctionnement du site, on notera que deux ingénieurs travaillent quotidiennement sur le projet pour l'équivalent de deux temps plein. L'activité principale du site étant le calcul LHC, un physicien pour chacun des groupes LHC du laboratoire fait le lien entre le site et les besoins des groupes ALICE et ATLAS.

2.5 La gestion

Les évolutions du site sont discutées dans le Comité Technique Informatique du laboratoire dirigé par un représentant du Conseil d'Unité et constitué par le personnel du service informatique et des représentants des activités scientifiques et techniques du laboratoire utilisant les ressources informatiques.

2.6 Financement

L'infrastructure qui accueille le site a été financée par le LPSC et par l'IN2P3. Les dépenses liées à sa consommation électrique sont supportées par le LPSC. Quant au matériel (serveurs de calcul et de stockage), il a été financé par les groupes de physique et le LPSC puis le site a reçu le support de l'Institut des Grilles et de LCG France.

LCG France assure le renouvellement du matériel après une période de 4 ans (5 ans à partir de 2012) mais avec un seuil minimal qui varie en fonction des années et des ressources dont il dispose. L'Institut des Grilles a financé le site en 2008 pour aider à son démarrage puis en 2009 pour favoriser l'implantation d'un site de type EGI en région grenobloise. Enfin en 2010, le LPSC et CIMENT, le « mésocentre [7] » des universités grenobloises, ont sollicité un financement commun auprès de l'IdG afin de débuter leur collaboration et d'augmenter la capacité de leurs sites respectifs en termes de stockage. Le budget de fonctionnement, personnel compris, est d'environ 100 000 € par an. Les ressources pour l'équipement ont varié de 100 000 à 200 000 € par an en fonction des années et diminuent en 2012 où environ 65 000 € ont été investi dans le matériel.

En 2012, pour la première fois, LCG-France n'a pas eu le budget suffisant pour subvenir au remplacement du matériel vieux de 4 ans dans les T2s français. Le renouvellement total du matériel ancien pour les années qui viennent n'est pas non plus assuré. Il est possible de rallonger quelque peu la durée d'utilisation du matériel mais cela ne permet que de reporter les dépenses. Des financements hors IN2P3, doivent donc être trouvés pour continuer d'accompagner les demandes des expériences.

2.7 Les activités du site

Le site du LPSC est principalement utilisé pour le calcul LHC. Les premières « Virtual Organisations » (VO) à y être installées ont donc été, outre les VO Ops et dteam permettant

7. Un mésocentre est un centre à portée régionale regroupant un ensemble de ressources informatiques et de moyens humains à destination d'une ou plusieurs communautés scientifiques et issus de plusieurs entités (EPST, Universités,Industriels, ...).

de tester le bon fonctionnement du site, les VO ATLAS et ALICE. Depuis, le site s'est ouvert à d'autres disciplines, qu'elles soient régionales (VO Rhônes-Alpes, EUMED pour les pays méditerranéens), ou qu'elles correspondent à des besoins d'équipes du LPSC (VO LPSC ouverte à tous les membres du LPSC, VO Calice pour le groupe participant à ILC, VO MURE dédiée à la simulation de réacteurs nucléaires, VO Biomed pour l'imagerie médicale). On notera d'autre part, qu'une collaboration s'est engagée autour des grilles de calcul et de l'informatique verte entre le service informatique du LPSC et le groupement de laboratoire pour le calcul intensif CIMENT. CIMENT vise au développement de projets de calcul de type mésocentre au sein des universités grenobloises. D'autre part, CIMENT a déployé une grille de calcul exploitée via le logiciel CIGRI qui permet de fédérer plusieurs clusters. L'objectif principal de la collaboration entre le LPSC et CIMENT est de donner accès aux utilisateurs de CIMENT à la grille EGI et réciproquement.

L'évolution de l'utilisation du site en termes de CPU depuis sa création est montrée sur la figure 12. Comme il a été déjà mentionné, la figure indique clairement que le site est majoritairement utilisé par les expériences LHC : ATLAS et ALICE.

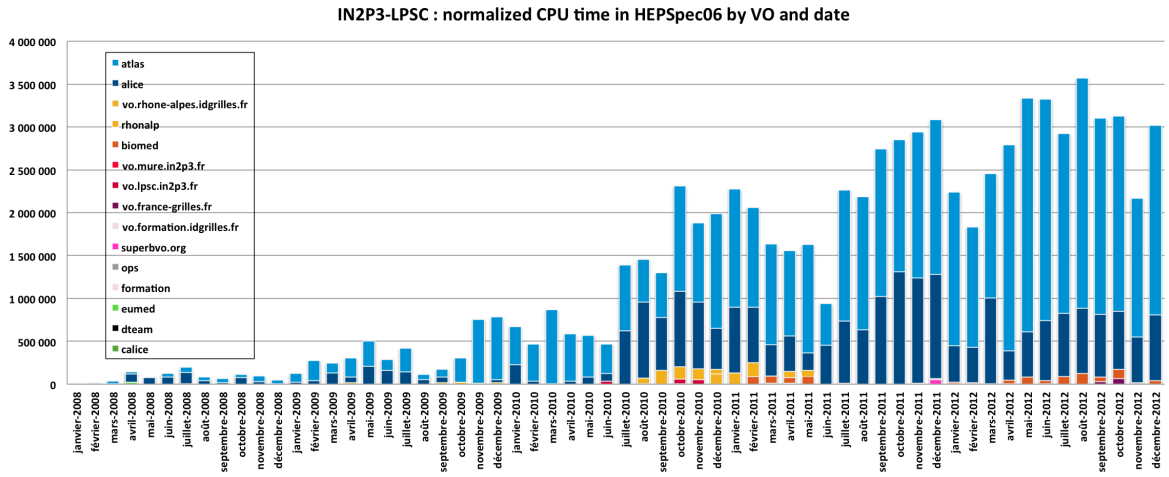


FIGURE 12 – Évolution du temps de calcul normalisé utilisé par les différentes VO supportés par le LPSC depuis la mise en production du site.

2.7.1 Les activités du site pour la collaboration ATLAS

Depuis la mise en fonctionnement du site, son utilisation par ATLAS a été continue (voir figure 12). L'activité du site du LPSC représente en 2012 environ 1 % de l'activité de l'ensemble des T2s et 10 % de celle des T2s français hors CC-IN2P3. 600 tâches ATLAS tournent en moyenne en continu sur le site (voir figure 13) ce qui représente environ 1 500 000 tâches sur l'année soit 2900 ans HEP-SPEC06.

Environ 60 % des tâches traitées, comme sur l'ensemble des sites ATLAS, sont des tâches d'analyse, les 40 % restant se partagent pour moitié en tâches de production et de test. La répartition en terme de CPU utilisé est très différente : l'analyse représente 25 % du CPU consommé, les tests 3 % et la production 70 %. Ces pourcentages sont similaires à ceux de la moyenne des T2s d'ATLAS sauf en ce qui concerne les tâches de reconstruction qui

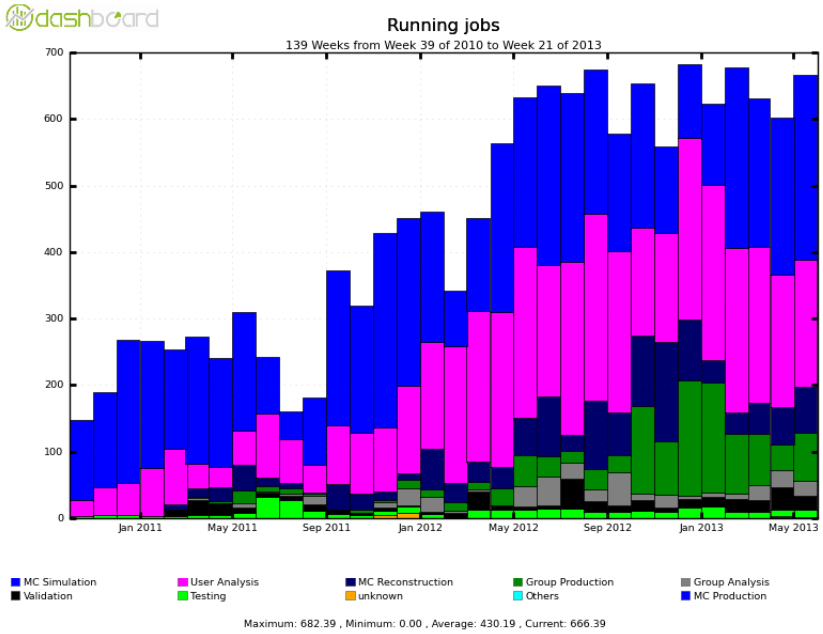


FIGURE 13 – Évolution des tâches ATLAS en cours de traitement en fonction de leur type pour le site IN2P3-LPSC.

représente 15 % des tâches au LPSC et un peu moins en moyenne sur l'ensemble des sites car ces tâches exigeantes ne sont pas distribuées à tous les sites.

Côté stockage, environ 540 TO sont disponibles et utilisés par l'expérience. Dans l'espace réservé aux données distribuées par ATLAS (hors données temporaires liées aux tâches en traitement et données des utilisateurs), le LPSC étant classé comme un site fiable, il reçoit 2,6 % des données ATLAS dont une grande fraction de données primaires (voir figure 14). Chaque mois 100 TO (50 TO) sont transférés vers (depuis) le LPSC.

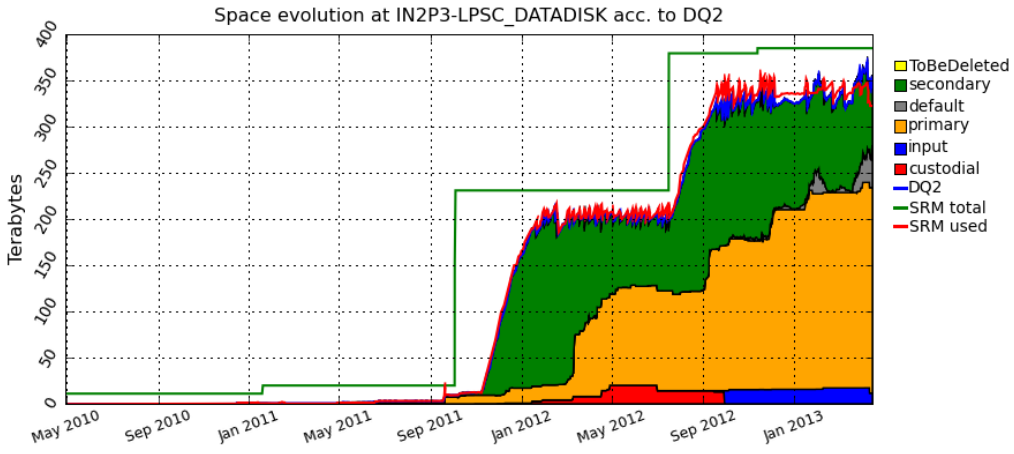


FIGURE 14 – Évolution des données stockées sur le site du LPSC (hors données temporaires liées aux tâches en traitement et données des utilisateurs).

Les performances du site sont très bonnes : sa disponibilité pour la production et l'analyse est de 95 % sur l'année 2012 et le pourcentage des tâches de production correctement

exécutées sur le site est supérieur à 90 %. Les erreurs les plus fréquentes sont liées au système de gestion des tâches et à l'accès aux données.

2.7.2 Les activités du site pour la collaboration ALICE

Selon le modèle de calcul de la Collaboration ALICE , l'ensemble des ressources des T2s et T3s est globalement dédié à la production Monte Carlo ainsi qu'à l'analyse des utilisateurs. Ainsi, depuis son déploiement, le site de Grenoble participe régulièrement aux campagnes de traitement des données d'ALICE.

En 2012, le site a traité en continu entre 150 et 300 tâches (voir figure 15), ce qui représente un total d'environ 6 000 000 de tâches soit environ 850 ans HEP-SPEC06. Cela correspond à 5,7 % des tâches traitées pour ALICE par les T2s français hors CC-IN2P3.

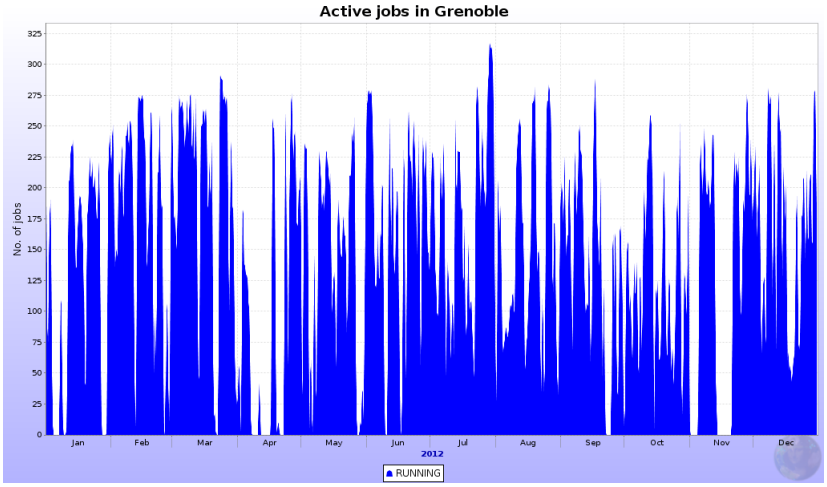


FIGURE 15 – Nombre de tâches en exécution pour ALICE au LPSC durant l'année 2012.

ALICE n'utilise que les espaces de stockage qui fonctionnent avec xrootd natif. Les données d'ALICE remplissent environ 85 % de cet espace soit 70 TO.

2.8 Performances

Lorsque le site du LPSC était un T3, il n'avait pas à s'assurer que le minimum de performances exigées pour les sites T2 soit atteint. Cependant, afin d'assurer un service de qualité à ses utilisateurs, le personnel du LPSC s'est attaché à maintenir le site à un très bon niveau de disponibilité et de fiabilité. Pour illustrer les performances du site depuis sa création, les valeurs de ces deux variables telles que définies et enregistrées par WLCG sont montrées en fonction du temps sur la figure 16. Elles sont supérieures à 90 % en moyenne. La baisse d'efficacité à la fin de l'automne 2012 est ponctuelle et liée au changement d'intergiciel (EMI).

Les performances relatives aux expériences sont bonnes comme décrit précédemment. L'équipe travaille à améliorer encore les résultats pour baisser les taux d'erreurs, en particulier pour ALICE, pour laquelle le suivi de l'expérience est moindre que pour ATLAS, qui bénéficie entre autre du travail des personnes du Squad.

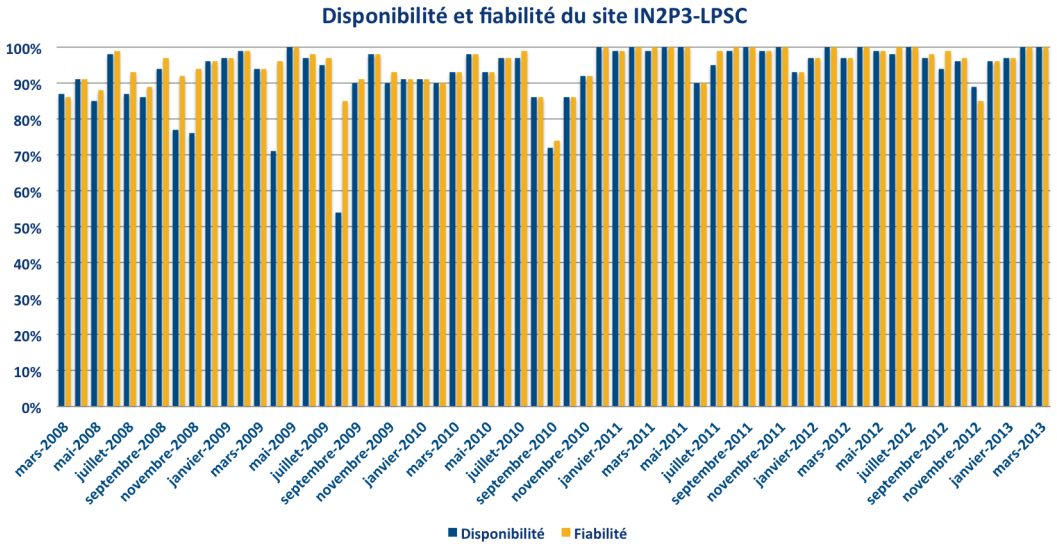


FIGURE 16 – Disponibilité (en bleu) et fiabilité (en jaune), d'après WLCG, du site du LPSC en fonction du temps.

3 Perspectives

Le site devra continuer d'accompagner les avancées des expériences LHC. Nous serons attentifs au suivi du raccordement du site à LHCONE car le réseau sera certainement le cœur des développements futurs. Les changements à court terme seront liés aux évolutions des expériences (calcul parallèle, changement de logiciel, accès aux données via WebDaV, xrootd pour les fédérations de stockage, ...) mais aussi au développement du site et des logiciels généraux (SL6).

Par ailleurs, il me semble intéressant de poursuivre et d'approfondir la collaboration qui a débuté entre CIMENT et le LPSC. Plusieurs applications utilisées pour des analyses de chercheurs du LPSC ont été portées sur la grille légère de CIMENT grâce à l'investissement d'un ingénieur de notre laboratoire. Ceci a permis de disposer pour ces seules analyses de ressources équivalentes à la moitié de celles du LPSC pendant le temps nécessaire à leur exécution. À l'inverse, la mise à disposition des ressources non utilisées du laboratoire pour la grille de CIMENT semble moins urgente actuellement, d'autant que le taux d'occupation du site du LPSC est important et que CIMENT vient de recevoir un nouveau cluster HPC (calcul intensif) qui doublera sa capacité de calcul. L'utilisation éventuelle de ce dispositif de calcul intensif par les membres du LPSC devra être étudiée.

En outre, cette collaboration avec les membres de CIMENT a permis aussi de définir un projet d'implantation d'une plateforme d'hébergement mutualisée avec l'Université Joseph Fourier pour accueillir au LPSC les ressources informatiques des laboratoires grenoblois. Ce projet permettra de développer les technologies vertes qui sont un pôle d'intérêt commun de l'ensemble des parties impliquées. L'implantation de cette nouvelle structure permettra certainement de renforcer la collaboration entre les différents sites informatiques grenoblois, de valoriser les investissements tournés vers l'informatique « verte » en particulier EcoClim au LPSC et ouvre la possibilité au laboratoire, s'il le souhaite, de développer ses ressources informatiques.

Troisième partie

Mesure de la section efficace de production de top-antitop

De part sa masse importante, le quark top tient une place particulière parmi les particules élémentaires aujourd'hui. Du point de vue expérimental d'abord : avec ses 173 GeV [39], il n'a pu être directement observé qu'à partir de la mise en service du Tevatron avec des collisions protons-antiprotons à 1,8 TeV. Il est aussi le seul quark dont le temps de désintégration est plus court que le temps nécessaire au processus d'hadronisation, ce qui rend sa signature expérimentale particulière et ouvre une fenêtre d'observation unique sur le comportement intrinsèque des quarks. Ceci permet par exemple d'étudier le spin du top, puisque ses effets sont directement transmis aux particules en lesquelles le top se désintègre sans être pollués par les effets non perturbatifs de l'hadronisation. Du point de vue théorique ensuite : la masse du top est proche de l'échelle d'énergie de la brisure électrofaible ce qui laisse penser qu'il peut jouer un rôle particulier dans les nouvelles théories actuellement explorées. En effet, l'échelle de Fermi qui caractérise cette brisure vaut $\mathrm{v} = (\sqrt{2}\mathrm{G_F})^{-\frac{1}{2}} = 246\ \mathrm{GeV}$, ce qui donne un couplage de Yukawa du top au Higgs de l'ordre de 1 ($\frac{\text{masse(top)}}{\frac{\mathrm{v}}{\sqrt{2}}}$). Le quark top est donc le seul fermion à posséder une masse naturelle dans le sens où elle est du même ordre de grandeur que l'échelle d'énergie du mécanisme qui l'engendre. D'autre part, plusieurs théories prédisent l'existence de nouvelles particules de grande masse qui pourraient se désintégrer préférentiellement en quarks top et à l'inverse le quark top pourrait se désintégrer en particules n'appartenant pas au modèle standard.

Le top est donc une signature intéressante pour tester la validité du modèle standard et sonder la nature en quête de nouveaux phénomènes. Dans ce cadre, après l'observation du quark top, la première analyse à mener consiste à mesurer avec précision sa section efficace de production dans les différents modes de production et de désintégration possibles et de vérifier la compatibilité des résultats avec ceux calculés dans le Modèle Standard. L'analyse décrite ici entre dans ce cadre. Il s'agit de mesurer la section efficace de production top-antitop ($\mathrm{t\bar{t}}$) dans le canal contenant un lepton et des jets avec les premières données du Run II du Tevatron et l'expérience DØ.

4 Le contexte expérimental

4.1 Le Tevatron et l'expérience DØ

Le Tevatron est le collisionneur proton-antiproton du Fermilab situé dans la périphérie de Chicago aux États-Unis. Il a connu deux grandes périodes de prise de données : le Run I de 1992 à 1996 pendant lequel il a fonctionné à une énergie dans le centre de masse de 1,8 TeV et le Run II de 2001 à 2011. Le Run I a été marqué par la découverte du quark top en 1995 [40]. Pendant le Run II, le Tevatron a fonctionné avec une énergie dans le centre de masse de 1,96 TeV ce qui a augmenté la section efficace de production de paires $\mathrm{t\bar{t}}$ de 40 % environ.

L'expérience DØ possède, avec CDF, l'un des deux détecteurs enregistrant les collisions

du Tevatron. Il s'agit d'un détecteur de particules de forme cylindrique, typique de ceux qu'on trouve auprès des collisionneurs. Il possède un calorimètre à échantillonnage à argon liquide et uranium et un spectromètre à muons dont la couverture a été étendue pour le Run II. On leur a ajouté un trajectographe composé d'un détecteur à micropistes de silicium (SMT) et d'un détecteur à fibres scintillantes (CFT), tous deux placés dans un solénoïde supraconducteur de 2 T, lui même situé à l'intérieur du calorimètre.

Le SMT a été conçu pour reconstruire les traces et les vertex des particules chargées jusqu'à une pseudorapidité de $|\eta| < 2,5$ [41]. Le calorimètre est divisé en 3 parties : une partie centrale (CC) qui couvre la région en rapidité jusqu'à $|\eta| \approx 1$ et deux bouchons (EC) qui étendent cette couverture jusqu'à $|\eta| \approx 4$, chacun étant contenu dans des cryostats séparés. Des scintillateurs, placés entre les calorimètres, permettent de couvrir les régions situées à $1,1 < |\eta| < 1,4$ [42]. Enfin, le détecteur de muons, situé autour des calorimètres, couvre la région allant jusqu'à $|\eta| < 2$. Il est composé de 3 couches de détecteurs et de scintillateurs servant au système de déclenchement. Les deux couches les plus extérieures sont placées autour d'un aimant toroïdal de 1,8 T [43].

4.2 Les données utilisées

Les données utilisées pour cette analyse ont été enregistrées par le détecteur DØ entre août 2002 et août 2004, c'est à dire pendant les premières années du Run IIa du Tevatron. Elles correspondent à une luminosité intégrée d'un peu plus de 420 pb^{-1}.

4.3 La production et la désintégration des quarks top

Le processus dominant de production de quarks top au Tevatron (et au LHC) est la production de paires top-antitop via l'interaction forte. Les diagrammes de production au premier ordre sont donnés sur la figure 17 ; l'annihilation quark-antiquark est le processus majoritaire au Tevatron, il correspond à 85 % de la section efficace de production. Son calcul pour une masse du quark top de 175 GeV/c^2 à l'ordre NNLO+NNNLL donne $6,77 \pm 0,42$ pb [44].

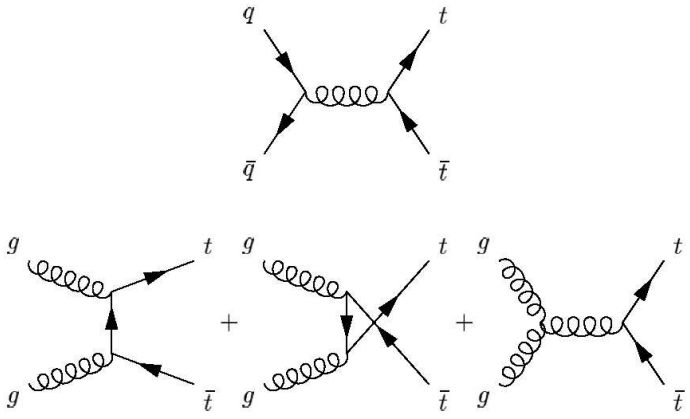


FIGURE 17 – Graphes de production au premier ordre de paires de quarks top par interaction forte.

Dans le cadre du modèle standard, le quark top se désintègre dans quasiment 100 % des cas en un boson W et un quark b : $\mathrm{BR(t \to Wb)} = |\mathrm{V_{tb}}|^2 = 99,9\ \%$ à 90 % de niveau

de confiance [45]. La signature des paires de quarks top est donc fonction du mode de décroissance des bosons W. On distinguera les canaux tout hadronique, leptonique et semi-leptonique dans lesquels respectivement les deux W se désintègrent en paire de quarks, en lepton et neutrino, un W en paire de quarks et l'autre en lepton et neutrino. C'est ce dernier canal qui a été choisi pour l'analyse décrite ici car il combine un rapport de branchement relativement grand ($14,8\ \%$ pour chaque type de lepton) et la présence d'un lepton isolé, ce qui permet de limiter le bruit de fond multi-jets à un niveau raisonnable.

5 L'analyse

5.1 Principe

L'objectif de l'analyse est donc de mesurer la section efficace de production de paires de top se désintégrant dans le canal semi-leptonique appelé aussi lepton+jets. L'état final, représenté sur la figure 18 (dans le cas où le lepton est un muon), contient donc :

- un lepton isolé de grande impulsion transverse et de l'énergie manquante provenant du neutrino qui échappe à la détection, ces deux particules étant issues d'un des W ;
- de 4 jets provenant de 2 quarks b issus de la désintégration des tops et de 2 quarks légers issus de celle du second W.

Les leptons tau sont plus difficiles à identifier et des développements seraient nécessaires pour les inclure complètement dans l'analyse. Ainsi seuls les cas où le lepton est un muon ou un électron seront considérés. Cependant les cas où le lepton tau se désintègre en électron ou muon sélectionné par l'analyse seront pris en compte.

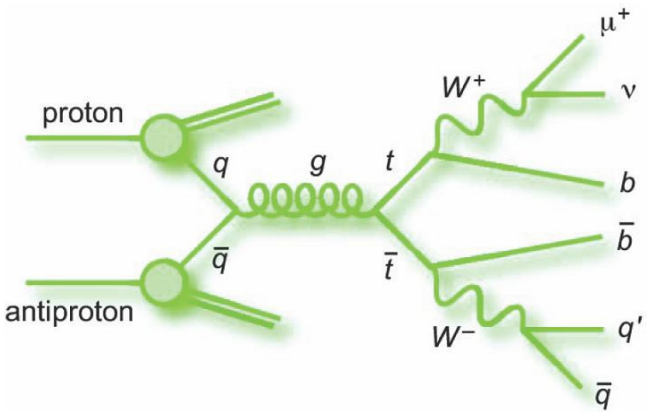


FIGURE 18 – Diagramme de production de paires de top dans le canal semi-leptonique à partir d'une collision proton-antiproton.

En conséquence, les principaux bruits de fond associés à cet état final seront

- la production de jets associés à un W (notée W+jets) qui présente le même état final que le signal mis à part la présence de quarks b liée à la désintégration des tops ;
- la production multiple de jets si un des jets a été mal reconstruit et identifié comme un lepton, ce qui induit aussi de l'énergie transverse manquante ;
- la production de jets associés à un Z se désintégrant en deux muons, dans les cas où un des muons se superpose à un jet et l'autre est identifié comme le muon isolé ;

– la production de jets associés à un Z se désintégrant en deux taus où la présence d'un lepton et d'énergie manquante provient de la désintégration des taus ;
– les évènements di-bosons ;
– les productions de paires de top dans le canal purement leptonique, dans ce cas on utilisera pour ces évènements la même section efficace que celle du canal semi-leptonique que l'on cherche à mesurer ;
– la production de quark célibataire via l'interaction faible.

Pour effectuer la mesure de la section efficace de paires de top, il s'agit donc, dans un premier temps, de sélectionner des évènements de type W+jets et ensuite de distinguer les évènements provenant de désintégrations de paires de top de ceux provenant de la production associée de jets avec un W. Jusqu'alors deux méthodes avaient été utilisées pour faire cette distinction : une analyse topologique et une analyse utilisant l'étiquetage par durée de vie des hadrons beaux. La première méthode utilise la topologie particulière des évènements $\mathrm{t\bar{t}}$ qui est due à la grande masse des quarks top produits quasiment au repos au Tevatron. La seconde repose sur l'identification des 2 quarks b issus de la désintégration des 2 tops via l'utilisation de la mesure de la distance entre le vertex primaire de la collision et les traces des hadrons B. En effet, ces hadrons ayant une durée de vie finie, leur désintégration définit un vertex différent du vertex primaire. Pour identifier les quarks b issus de la désintégration des quarks top, il est aussi possible d'utiliser une autre de leur propriété qui est leur assez grande probabilité de se désintégrer de façon semi-leptonique. C'est cette dernière méthode, complémentaire des deux précédentes, qui sera utilisée dans cette analyse.

En effet, si l'on considère qu'il y a 2 quarks b pour chaque évènement $\mathrm{t\bar{t}}$ et qu'en moyenne un quark c est produit dans une désintégration sur 3 d'un boson W, si on prend en compte de plus les rapports de branchement suivants [45] :

- $\mathrm{b} \rightarrow \mu = 11.0\ \%$
- $\mathrm{b} \rightarrow \mathrm{c} \rightarrow \mu = 8.0\ \%$
- $\mathrm{c} \rightarrow \mu = 8.7\ \%$

alors environ $40\ \%$ des évènements $\mathrm{t\bar{t}}$ contiennent un muon de faible impulsion dans un jet. On notera que ce pourcentage serait le même si on cherchait à mettre en évidence un électron dans un jet. Cependant électrons et jets sont détectés par les mêmes appareils, les calorimètres, ce qui rend l'identification de l'électron non isolé plus difficile.

Une fois les évènements sélectionnés et étiquetés par la présence d'un muon dans un jet, il faut évaluer le nombre d'évènements de bruits de fond restant. Les sections efficaces de production de la plupart des bruits de fond considérés ont la particularité de diminuer en fonction du nombre de jets associés, ceux-ci provenant généralement de radiations de gluons. L'analyse sera donc conduite en fonction du nombre de jets reconstruits (1, 2, 3 et 4 ou plus). Les cas où il y a 1 ou 2 jets reconstruits correspondent à une configuration où les bruits de fond sont dominants, ils seront donc utilisés pour s'assurer que ces derniers sont sous contrôle. Les cas où il y a 3 ou 4 et plus jets reconstruits seront utilisés pour la mesure de la section efficace.

Connaissant le nombre d'évènements sélectionnés dans les données, $\mathrm{N_{données}}$, et l'évaluation du nombre d'évènements de bruits de fond $\mathrm{N_{bruit\ de\ fond}}$, la section efficace $\sigma(\mathrm{t\bar{t}})$ est donnée par l'équation suivante :

$$\sigma(\mathrm{t\bar{t}}) \;=\; \frac{\mathrm{N}_{\mathrm{t\bar{t}}}}{\varepsilon \times \mathrm{Br} \times \mathcal{L}} \;=\; \frac{\mathrm{N}_{\text{données}} \;-\; \mathrm{N}_{\text{bruit de fond}}}{\varepsilon \times \mathrm{Br} \times \mathcal{L}} \tag{1}$$

où ε est le produit de l'acceptance et de l'efficacité de sélection des évènements $\mathrm{t\bar{t}}$ dans le canal lepton+jets. ε est mesuré sur des évènements $\mathrm{t\bar{t}}$ simulés par Monte Carlo et corrigé pour d'éventuelles différences entre les données réelles et la simulation. Br est le rapport de branchement correspondant au canal lepton+jets et $\mathcal{L}$ est la luminosité intégrée.

Cette mesure est faite dans les 4 canaux considérés : canaux où le lepton est un électron (noté e+jets) ou un muon (noté μ+jets), et pour lesquels 3 ou 4 jets et plus ont été reconstruits. Pour prendre en compte au mieux toute l'information de chacun de ces canaux, la section efficace sera obtenue par le calcul d'un maximum de vraisemblance.

Les principales étapes de cette analyse sont décrites dans les paragraphes qui suivent, le détail de l'analyse se trouve dans une note de l'expérience DØ [46] et dans la thèse de Florent Chevallier [47]. D'autre part, l'analyse a été approuvée et rendue publique par la collaboration DØ [48].

5.2 Identification et reconstruction des objets

Les évènements tels que produits par les collisions du Tevatron et enregistrés par les détecteurs ainsi que les données issues de la simulation Monte Carlo sont traitées par le programme DØreco [49], élaboré par la collaboration pour reconstruire les objets physiques associés. L'identification et la reconstruction des objets utilisés dans cette analyse sont brièvement décrits dans les paragraphes suivants.

5.2.1 La reconstruction du vertex primaire

L'algorithme DØreco est utilisé pour reconstruire le vertex de la collision nécessaire à la mesure des impulsions des objets. Il est basé sur une technique d'ajustement en deux passages qui minimise le paramètre d'impact des traces (de qualité croissante pour chaque passage) par rapport aux vertex supposés.

Le choix du vertex primaire parmi les vertex reconstruits est basé sur la distribution en impulsion transverse des traces associées : s'il s'agit d'un processus dur, les traces associées au vertex primaire auront en moyenne une impulsion transverse plus élevée. La description et les performances de l'algorithme sont présentées dans la référence [50].

5.2.2 Les électrons

L'algorithme de reconstruction des objets électromagnétiques est basé sur un algorithme de cone simple qui rassemble en un amas les cellules du calorimètres se trouvant à une distance angulaire ΔR[8] inférieure à 0,2 de cellules germes (c'est-à-dire des cellules possédant une énergie transverse supérieure à 15 GeV). Différentes propriétés sont ensuite mesurées pour chaque amas tel que son centre de gravité, ses dimensions, l'énergie électromagnétique et l'énergie totale (électromagnétique et hadronique) qu'il contient. À partir de ces variables,

8. $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ où ϕ est l'angle azimutal et η la pseudo-rapidité.

des critères de qualité sont définis comme la fraction électromagnétique de l'amas, l'isolation, les caractéristiques géométriques de la gerbe, des critères d'association à une trace, ainsi que la distance au vertex de la trace associée.

5.2.3 Les muons

Les muons sont reconstruits en utilisant les informations données par les chambres à muons et par les détecteurs de traces. La qualité de la reconstruction du muon est déterminée en fonction du nombre de couches touchées du détecteur à muons ainsi que de la qualité de l'ajustement de la trace dans ce détecteur. Pour améliorer la résolution en impulsion, la trace reconstruite par le système à muons est associée à une trace du trajectographe central et un ajustement global est effectué.

5.2.4 Les jets

Les jets sont reconstruits à partir des amas d'énergie de plus de 8 GeV dans le calorimètre et après un nettoyage des cellules isolées de basse énergie et des cellules avec un assez grand taux de bruit. L'algorithme utilisé pour reconstruire les jets est appelé « improve legacy cone algorithm » [51], il est utilisé avec une taille de cône égale à 0,5. Les jets utilisés dans l'analyse doivent en outre passer des critères de qualité dépendant de leur fraction d'énergie électromagnétique, des rapports en énergie entre les cellules et de leur association avec un jet détecté par le système de déclenchement de niveau 1.

5.2.5 L'énergie transverse manquante

L'impulsion transverse des partons à l'origine des collisions étudiées étant très faible, la présence de particules comme le neutrino qui n'interagissent pas ou peu avec le détecteur peut être détectée en mesurant un déséquilibre dans l'énergie mesurée dans le plan transverse. L'énergie transverse manquante ($\not\!\!E_{\mathrm{T}}$) est reconstruite comme l'opposée de la somme vectorielle des cellules (considérées comme non bruyantes) des calorimètres auxquelles sont ajoutés les muons.

5.2.6 Étalonnage et ajustement du Monte Carlo

Les mesures obtenues sur les objets décrits ci-dessus font bien entendu l'objet d'un étalonnage précis qu'il serait trop long de détailler dans ce document. De même, les différences d'étalonnage et de résolutions entre les données réelles et la simulation sont étudiées et la simulation est ajustée si besoin pour que la représentation des différentes variables utilisées dans l'analyse soit conforme à la réalité.

5.3 Présélection des évènements

Il s'agit donc de sélectionner des collisions ayant produit un lepton (électron ou muon) isolé et de grande impulsion transverse, de l'énergie transverse manquante et un ou plusieurs jets. La même sélection sera appliquée aux données et à la simulation. Mis à part la sélection du lepton et les coupures sur l'énergie transverse manquante, les mêmes critères de sélection sont utilisés pour les canaux électron et muon.

5.3.1 Sélection du système de déclenchement

La première étape de la sélection se fait au niveau du système de déclenchement de l'expérience. À l'issue de cette étape, les données utilisées par l'analyse sont réparties en deux groupes contenant soit au moins un électron soit au moins un muon d'impulsion transverse supérieure à 15 GeV/c et, dans les deux cas, un jet d'énergie transverse supérieure à 20 GeV selon la classification du système de déclenchement. Plusieurs combinaisons de conditions du système de déclenchement sont utilisés pour sélectionner les données utilisées, la luminosité intégrée correspondante est respectivement de 422 ± 26 et 426 ± 26 pb^{-1} pour le canal μ+jets et e+jets.

Pour les évènements issus de la simulation, on attribue à chacun d'entre eux un poids correspondant à la probabilité que l'évènement en question remplissent les conditions imposées par le système de déclenchement. Cette probabilité est calculée à partir de la mesure de l'efficacité de sélection de chaque niveau du système de déclenchement. Cette méthode permet d'utiliser l'ensemble des évènements simulés et donc de diminuer l'incertitude statistique associée.

5.3.2 Présélection lâche

Les évènements correspondant à une paire de quarks top qui se désintègrent dans les canaux contenant un lepton et des jets sont caractérisés par 2 jets de quark b, un lepton isolé de grande impulsion transverse et un neutrino provenant de la désintégration du W dit leptonique et deux jets légers provenant de la désintégration du W dit hadronique. D'autres jets provenant de rayonnements des états initial et final peuvent aussi être présents. Les données sont donc sélectionnées en imposant les conditions suivantes :

- un électron ou un muon isolé avec une impulsion transverse supérieure à 20 GeV/c dans le calorimètre central ($|\eta| < 1,1$) pour l'électron et avec $|\eta| < 2,0$ pour le muon. L'électron est dit isolé si $\frac{\mathrm{E(0,4)-E(0,2)}}{\mathrm{E(0,2)}} < 0,15$ où $\mathrm{E(R)}$ est l'énergie contenue dans un cône de rayon R dans le plan $\eta - \phi$ ($\mathrm{R} = \sqrt{\eta^2 + \phi^2}$). Un muon est considéré comme isolé si sa distance angulaire au jet le plus proche est supérieure à 0,5.
- L'énergie transverse manquante $\not\!\!E_T$ doit être supérieure à 20 GeV et ne doit pas être colinéaire à la direction du lepton sélectionné dans le plan transverse, ce qui pourrait être caractéristique d'évènements de bruit de fond multi-jets mal reconstruits.
- Le jet de plus grande impulsion doit avoir une impulsion transverse supérieure à 40 GeV/c et les autres jets une impulsion transverse supérieure à 20 GeV/c.
- Les évènements avec un deuxième lepton de grande impulsion transverse (supérieure à 15 GeV/c) sont rejetés pour ne pas sélectionner d'évènements pris en compte par les analyses du canal purement leptonique.

Ces premières conditions définissent une présélection dite lâche des évènements qui sera utilisée par la suite pour évaluer les bruits de fond résiduels. Elle ne permet cependant pas d'éliminer une grande partie des bruits de fond W+jets et multi-jets : une sélection plus stricte est nécessaire. Cette sélection stricte est basée sur des coupures sur la qualité de l'électron et l'isolation des muons.

5.3.3 Présélection stricte

Dans le cas du canal muon, le bruit de fond principal aux désintégrations du W en muon provient de jets de saveur lourde se désintégrant de façon semi-leptonique. Les muons issus de ces désintégrations tendent à être non isolés et d'assez bas moment transverse. La coupure stricte porte donc sur l'isolation du muon :

$$\mathrm{CaloE}(0,1;0,4)/\mathrm{p_{T}}_{\mu} \; < \; 0,08 \quad \textbf{et} \quad \mathrm{TrkCone}(0,5)/\mathrm{p_{T}}_{\mu} \; < \; 0,06$$

où $\mathrm{p_{T}}_{\mu}$ est l'impulsion transverse du muon, $\mathrm{CaloE(r1;r2)}$ est l'énergie calorimétrique mesurée dans un cône creux de rayon interne r1 et externe r2 centré sur l'axe du muon et où $\mathrm{TrkCone(r)}$ est la somme de l'impulsion des traces se trouvant dans un cône de rayon r autour du muon à l'exclusion de l'impulsion de ce dernier.

Dans le cas du canal électron, le bruit de fond provient essentiellement de jets identifiés comme des électrons. Pour mieux séparer les électrons des jets, on construit un maximum de vraisemblance [52] basé sur plusieurs variables discriminantes telles que la fraction d'énergie électromagnétique, des variables mesurant la forme de la gerbe dans les calorimètres, le rapport entre l'énergie du calorimètre et l'impulsion de la trace associée, la qualité de l'association de l'amas du calorimètre avec la trace, la distance au vertex primaire, des variables d'isolation, etc.

5.3.4 Efficacité de la présélection pour le signal

Au final, l'efficacité de la présélection complète (après la coupure stricte) pour les multiplicités en jets supérieures ou égales à 3 jets est de 21 % et 22 % respectivement dans les canaux muon et électron.

5.4 Étiquetage des quarks b

Environ 40 % des évènements $\mathrm{t\bar{t}}$ présentent un muon dans un jet. Identifier un tel muon permet de réduire significativement les bruits de fond de l'analyse, car pour la plupart d'entre eux, la présence d'un quark b ou c associé ne peut provenir que d'un gluon rayonné, ce qui n'arrive qu'avec une probabilité relativement faible. De plus, les seules façons de réduire le nombre d'évènements W+jets sont d'utiliser l'étiquetage des b et la topologie des évènements $\mathrm{t\bar{t}}$. Cette analyse basée sur la méthode d'étiquetage qui consiste à identifier un muon non isolé de basse impulsion est ainsi complémentaire de l'analyse topologique et de l'analyse utilisant l'étiquetage des b basée sur la durée de vie des hadrons beaux. L'étiquetage des b est effectué après la présélection décrite au paragraphe précédent et est la dernière opération de la sélection finale.

L'étiquetage des jets par la présence d'un muon non isolé a été développé sur des jets d'énergie supérieure à 15 GeV et avec $|\eta|$ inférieur à 2 pour rester dans l'acceptance du détecteur de muons. Un jet est considéré comme étiqueté si, à l'intérieur d'un cône de rayon 0,5 (dans le plan $\eta - \phi$) autour de l'axe du jet, un muon est identifié avec les propriétés suivantes :

- muon de qualité moyenne,
- avec $p_T > 4\ GeV/c$,

- et $|\eta| < 2,0$,
- l'association du muon à une trace doit être de bonne qualité.

Un évènement est considéré comme étiqueté si au moins un de ses jets (qui passent les critères de sélections) est étiqueté comme b selon les conditions précédentes.

Les corrections et incertitudes à prendre en compte dans l'utilisation de cet étiquetage sont décrits dans les paragraphes qui suivent ainsi que ses performances. Les détails de ces études peuvent être trouvés dans la référence [53] qui est malheureusement une note interne à la collaboration DØ .

5.4.1 Efficacité d'étiquetage

L'efficacité d'étiquetage d'un jet de b contenant un muon a été mesurée sur des évènements $\mathrm{t\bar{t}}$ et multijets $b\bar{b}$ simulés par Monte Carlo et vaut respectivement $46,6 \pm 0,2\ \%$ et $12,3 \pm 2,2\ \%$.

Il est difficile de paramétrer cette efficacité en fonction de l'énergie transverse des jets de façon indépendante du processus physique étudié. En effet, la mesure de l'énergie transverse elle-même est corrélée à la présence d'un muon dans le jet, puisque l'énergie du jet mesurée par le calorimètre est corrigée en fonction du $\mathrm{p_T}$ du muon en question lorsqu'il est présent. De plus la probabilité d'étiquetage est sensible à la présence de gluon allant en 2 b qui ont d'autant plus de chance d'être étiqueté qu'il y a 2 muons présents. Ceci implique que l'efficacité d'étiquetage dépend de l'énergie du jet mais aussi du processus étudié et doit donc être mesurée pour chacun d'entre eux.

5.4.2 Facteur correctif entre données et simulation

Les efficacités d'étiquetages pour l'analyse seront évaluées à partir de la simulation, les éventuelles différences entre les données et la simulation doivent donc être étudiées et prises en compte. Ces différences sont mesurées en utilisant des évènements de la résonance $J/\psi \to \mu^+\mu^-$ sélectionnés dans les données et simulés par Monte Carlo. La cinématique des muons étant bien reproduite par la simulation, la simulation ne doit être corrigée que par un facteur d'échelle de $0,945 \pm 0,024$ qui prend en compte l'identification du muon, de la trace associée et leur efficacité d'association.

5.4.3 Taux d'étiquetage des jets légers

Le taux d'étiquetage des jets légers doit aussi être pris en compte. Il est mesuré sur des évènements di-jets et vaut respectivement (par jet) $0,188 \pm 0,065\ \%$ et $0,163 \pm 0,009\ \%$ sur les données et le Monte Carlo. Un facteur correctif égal au rapport de ces deux taux sera donc appliqué aux jets légers simulés pour que les performances soient similaires à celles des données.

5.4.4 Efficacité d'étiquetage pour les évènements $\mathrm{t\bar{t}}$

L'efficacité d'étiquetage des évènements $\mathrm{t\bar{t}}$ dans le canal lepton+jets a été mesurée après la présélection stricte et après application des facteurs de correction. Pour le canal e+jets, elle vaut respectivement $15,7 \pm 0,3\ \%$ et $17,6 \pm 0,3\ \%$ pour une multiplicité en jets de 3 et 4

jets et plus. Ces efficacités sont de $15,5 \pm 0,4$ % et $16,9 \pm 0,4$ % pour le canal μ+jets.

L'efficacité de sélection globale (présélection stricte et étiquetage) est alors de 3,6 % et 3,3 % respectivement pour les canaux e+jets et μ+jets.

De la même façon, l'efficacité d'étiquetage des évènements $t\bar{t}$ dans le canal dilepton est de ($20,0 \pm 0,7$ % et $19,3 \pm 2,0$ %) pour le canal électron (multiplicité : 3 jets, 4 jets et plus) et ($18,3 \pm 0,7$ %, $20,3 \pm 2,2$ %) pour le canal muon respectivement. Ceci correspond à une efficacité totale de sélection de 0,8 % et 0,9 %.

5.4.5 Efficacité d'étiquetage des bruits de fond

L'efficacité d'étiquetage de chacun des bruits de fond est mesurée directement à partir de la simulation et après la sélection des évènements. Elle varie de 10 à 15 % pour les évènements où le top est produit en célibataire et peut atteindre quelques pour-cent pour les évènements de bruit de fond contenant un ou des bosons Z et W dans le canal muon (le taux d'étiquetage étant négligeable pour ces même bruits de fond dans le canal électron). L'efficacité d'étiquetage des évènements W+jets est de quelques pour-cent, voir table 1 .

TABLE 1 – Efficacité d'étiquetage des évènements W+jets dans la simulation (%).

Échantillon	1 jet	2 jets	3 jets	4 jets
W+jets	0,84±0,11	1,28±0,27	1,75±0,28	2,10±0,46

On ne mesure pas directement l'efficacité d'étiquetage des évènements du bruit de fond multi-jets mais on évalue sa contribution dans les données directement une fois l'étiquetage effectué.

Mesure alternative de l'efficacité d'étiquetage des évènements multi-jets et W+jets

Une méthode alternative peut être utilisée pour mesurer cette efficacité directement sur les données pour les évènements multi-jets et W+jets. Cette mesure, développée pour améliorer la précision sur l'efficacité d'étalonnage, ne sera utilisée ici qu'à titre de vérification des mesures obtenues à partir de la simulation pour les bruits de fond multi-jets et W+jets.

L'échantillon de données utilisé correspond à des critères de déclenchement demandant au moins 2 jets d'impulsion transverse supérieure à 10 GeV/c (éventuellement 5 GeV/c pour les jets suivants). Dans l'échantillon, au moins 3 jets doivent être sélectionnés avec les critères de l'analyse. On s'attend à ce que la probabilité d'étiquetage par jet varie en fonction de la position en ϕ et η du jet dans le détecteur ainsi que de son $\mathrm{p_T}$. C'est ce qui est montré sur le figure 19.

Il s'agit de paramétrer cette probabilité par jet (notée TRF pour « Tag Rate Function ») en fonction de la position reconstruite des jets η_{det} et ϕ_{det} et de leur impulsion transverse p_T. Pour cela, on suppose que cette dernière se factorise de la façon suivante :

$$\mathrm{TRF}(\mathrm{p_T}, \eta_{\mathrm{det}}, \phi_{\mathrm{det}}) \;=\; \mathrm{A}(\mathrm{p_T}, \eta_{\mathrm{det}}) \times \mathrm{B}(\phi_{\mathrm{det}}, \eta_{\mathrm{det}})$$

La fonction A est obtenue à partir d'une fonction analytique ajustée sur les données pour 100 tranches en η_{det}. A est obtenue à partir de la distribution en efficacité d'étiquetage en fonction du $\mathrm{p_T}$ du jet. La figure 20 montre cette distribution pour 3 tranches en η.

La fonction B représente la probabilité d'étiquetage en fonction de ϕ_{det}, elle est mesurée dans 5 intervalles en η tels que définis sur la figure 21. Chaque graphe est obtenu en divisant chaque valeur par la probabilité d'étiquetage moyenne.

Ceci permet un bon paramétrage de la probabilité d'étiquetage qui respecte la géométrie du détecteur à muons. On notera par exemple sur la distribution en ϕ de la figure 19, le creux dans la probabilité d'étiquetage lié à la présence des supports du détecteur de muons de l'expérience.

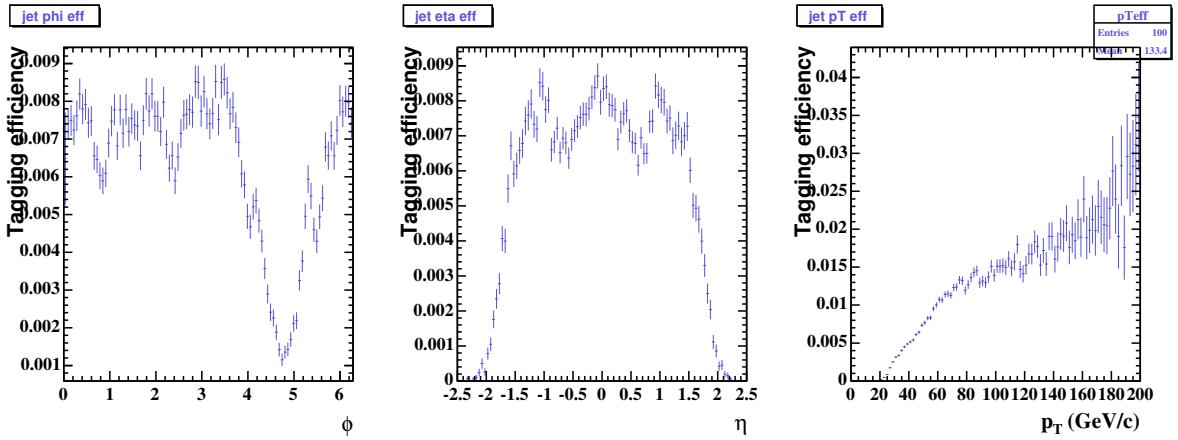


FIGURE 19 – Probabilité d'étiquetage des jets en fonction de l'angle reconstruit ϕ_{det} (à gauche), η_{det} (au centre) et $\mathrm{p_T}$ (à droite).

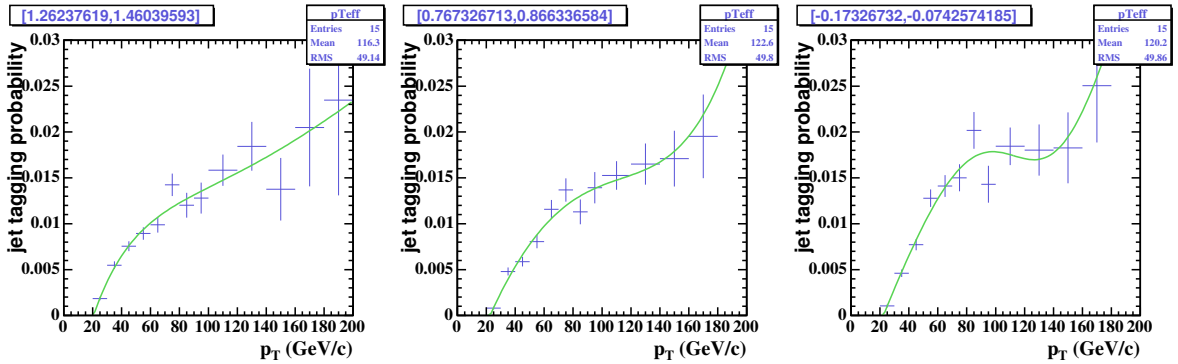


FIGURE 20 – Exemples de variation des TRF en fonction du p_T des jets pour 3 des 101 tranches en η_{det} : $1,26 < \eta_{det} < 1,46$, $0,77 < \eta_{det} < 0,87$ et $-0,17 < \eta_{det} < -0,07$.

Les probabilités d'étiquetage obtenues ainsi peuvent être appliquées aux échantillons W+jets en supposant que les bruits de fond multi-jets et W+jets ont la même composition en saveur. Ceci est vrai au premier ordre car la production de quarks lourds se fait, dans les deux cas, principalement via la production de deux quarks b par un gluon. La production de quark c associée au W induit une différence dans la composition en saveur des deux échantillons, mais cette différence reste inférieure à 10 % .

En faisant l'exercice on obtient une probabilité d'étiquetage des évènements W+jets de $1,99 \pm 0,04$ pour W+3 jets et $2,75 \pm 0,11$ pour W+>4 jets dans le canal électron. Ces résultats sont en accord avec ceux obtenus avec la simulation : $2,16 \pm 0,52\%$ et $2,25 \pm 0,86\%$ respectivement. Les résultats sont similaires dans le canal muon avec $2,06 \pm 0,04\%$ et $2,63 \pm 0,09\%$ pour l'étiquetage mesuré sur les données et $1,51 \pm 0,34\%$ et $2,04 \pm 0,55\%$ pour la simulation.

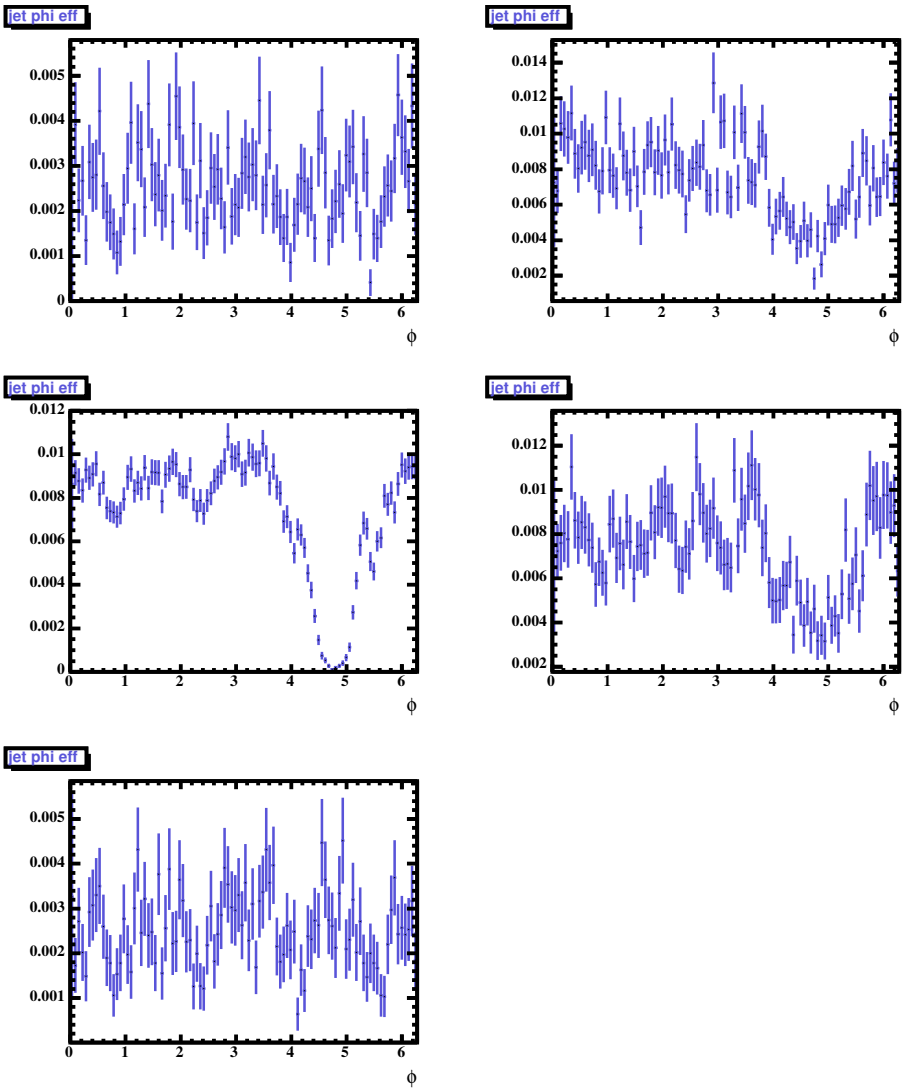


FIGURE 21 – Probabilité d'étiquetage des jets en fonction de ϕ_{det} dans 5 tranches en η_{det} : $\eta_{det} < -1,5$, $-1,5 < \eta_{det} < -1,0$, $-1,0 < \eta_{det} < 1,0$, $1,0 < \eta_{det} < 1,5$ et $\eta_{det} > 1,5$.

Cette méthode basée sur les données est donc prometteuse car elle présente des erreurs statistiques moindre que celle basée sur la simulation. Les incertitudes systématiques associées doivent cependant être évaluées avec soin avant de pouvoir l'utiliser.

5.5 Évaluation des bruits de fond

5.5.1 Méthode

Les bruits de fond di-bosons et top célibataires sont obtenus directement à partir de la simulation. Le bruit de fond Z+jets est évalué avec la simulation après avoir ajusté sur les données le nombre d'évènements Z+jets obtenus après la sélection lâche. On se base principalement sur les données pour mesurer le nombre d'évènements W+jets et multi-jets avec la méthode de la matrice. Celle-ci est appliquée deux fois : une première fois après étiquetage des b pour obtenir directement le nombre d'évènements multi-jets étiquetés et une deuxième fois avant l'étiquetage pour évaluer les W+jets. Le nombre de W+jets étiquetés est ensuite obtenu grâce aux probabilités d'étiquetage mesurées avec la simulation.

Méthode de la matrice

La méthode de la matrice utilise le fait qu'un évènement de bruit de fond n'a pas la même probabilité qu'un évènement de signal de remplir une condition stricte d'identification. Elle est décrite rapidement ci-après :

Si $\mathrm{N}_{\text{lâche}}$ est le nombre d'évènement sélectionné après un certain niveau de sélection lâche

$$\mathrm{N}_{\text{lâche}} = \mathrm{N}_{\text{signal}} + \mathrm{N}_{\text{bruit de fond}} \tag{2}$$

où $\mathrm{N}_{\text{signal}}$ et $\mathrm{N}_{\text{bruit de fond}}$ sont respectivement le nombre d'évènements de signal et de bruit de fond à ce niveau de sélection. On applique une coupure supplémentaire à la sélection sur une variable qui agit différemment sur le bruit de fond et le signal et on obtient $\mathrm{N}_{\text{stricte}}$ données sélectionnées qui vaut :

$$\mathrm{N}_{\text{stricte}} = \varepsilon_{\text{signal}}\, \mathrm{N}_{\text{signal}} + \varepsilon_{\text{bruit de fond}}\, \mathrm{N}_{\text{bruit de fond}} \tag{3}$$

où $\varepsilon_{\text{signal}}$ et $\varepsilon_{\text{bruit de fond}}$ sont respectivement la probabilité que le signal et le bruit de fond passent cette coupure. Si $\varepsilon_{\text{signal}}$ et $\varepsilon_{\text{bruit de fond}}$ sont mesurés par ailleurs, $\mathrm{N}_{\text{signal}}$ et $\mathrm{N}_{\text{bruit de fond}}$ sont déduits facilement des équations (2) et (3) :

$$\mathrm{N}_{\text{signal}} = \frac{\mathrm{N}_{\text{stricte}} - \varepsilon_{\text{bruit de fond}}\, \mathrm{N}_{\text{lhatache}}}{\varepsilon_{\text{signal}} - \varepsilon_{\text{bruit de fond}}} \quad \text{et} \quad \mathrm{N}_{\text{bruit de fond}} = \frac{\varepsilon_{\text{signal}}\, \mathrm{N}_{\text{lhatache}} - \mathrm{N}_{\text{stricte}}}{\varepsilon_{\text{signal}} - \varepsilon_{\text{bruit de fond}}} \tag{4}$$

5.5.2 Bruits de fond di-bosons et top

Les bruits de fond WW, WZ et ZZ sont évalués sur les échantillons Monte Carlo ainsi que la production électrofaible de quarks top.

5.5.3 Bruits de fond Z+jets

On considère ici que ALPGEN [54], le logiciel avec lequel les évènements Z+jets ont été générés, reproduit bien leur cinématique. Par contre, la section efficace de ces évènements pour chaque multiplicité en jet est corrigée à partir des données en mesurant le nombre d'évènements Z présélectionnés. Un facteur de correction K_Z est ainsi déduit pour prendre en compte les différences entre le Monte Carlo et les données.

Évaluation du facteur K_Z :
Pour chaque échantillon $Z \to \mu^+\mu^- + i$ jets produit avec ALPGEN, la masse invariante di-muons est reconstruite mais uniquement pour les évènements qui contiennent exactement i jets sélectionnés. Ceci permet dans une certaine mesure une association ad hoc entre les jets générés et reconstruits et évite de compter plusieurs fois le même type d'évènement dans les différents échantillons.

De la même façon, la masse invariante di-muons est reconstruite dans les données après exactement les mêmes coupures c'est-à-dire une fois la sélection du muon de haut p_T effectuée. La masse invariante est calculée à partir de ce muon et de tout autre muon identifié avec des critères plus lâches (associé à une trace et à une distance angulaire à plus de 0,5 d'un jet).

Les distributions obtenues sont ajustées à l'aide d'une Breit-Wigner normalisée convoluée à une Gaussienne pour le Z auxquelles on ajoute une exponentielle pour le bruit de fond. Le nombre d'évènements dans le pic du Z est obtenu directement avec les paramètres d'ajustement. La figure 22 illustre ceci pour le cas où le Z est accompagné de 3 jets. Le facteur K_Z est le rapport entre le nombre d'évènements dans le pic du Z pour les données et le Monte Carlo pour une luminosité intégrée donnée. La statistique n'étant pas suffisante pour évaluer K_Z dans le cas où il y a plus de 4 jets séparément, le facteur est mesuré dans le cas où il y a plus de 3 jets. Les mesures obtenues sont résumées dans la table 2.

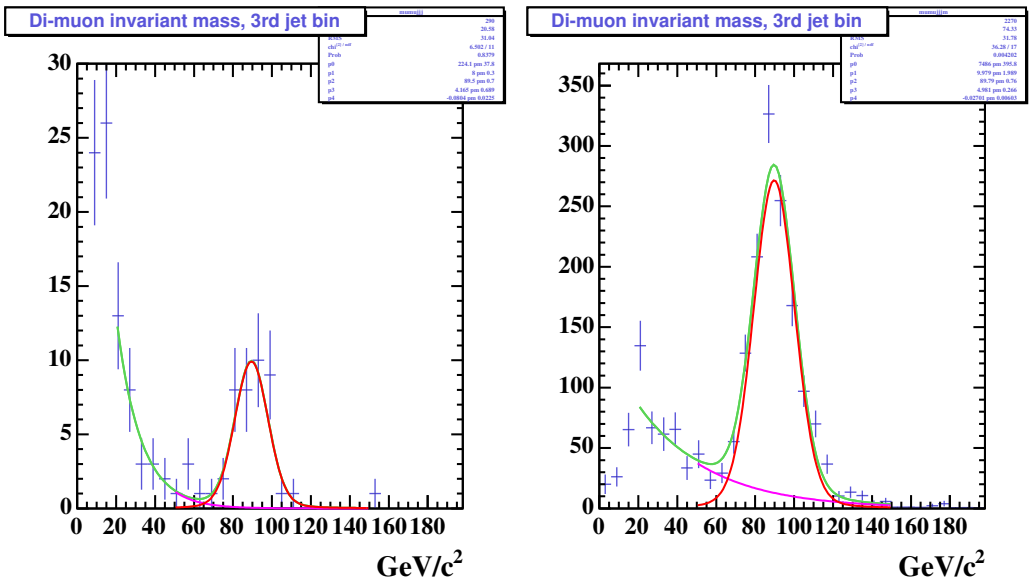


FIGURE 22 – Distribution de la masse invariante di-muons et ajustement des données (à gauche) et du Monte Carlo (à droite) pour une multiplicité en jets supérieure ou égale à 3.

TABLE 2 – Facteur K_Z en fonction de la multiplicité en jets.

multiplicité en jets	1	2	3 et plus
K_Z	1.42 ± 0.05	1.28 ± 0.09	0.97 ± 0.18

Afin de vérifier la validité de ces mesures, les facteurs obtenus sont appliqués au Monte

Carlo et les distributions obtenues sont comparées aux données, voir figure 23. De même, les distributions de la masse invariante di-muons à partir des données et du Monte Carlo corrigé sont comparées en retirant la coupure sur la masse du Z et montrent un bon accord (figure 24).

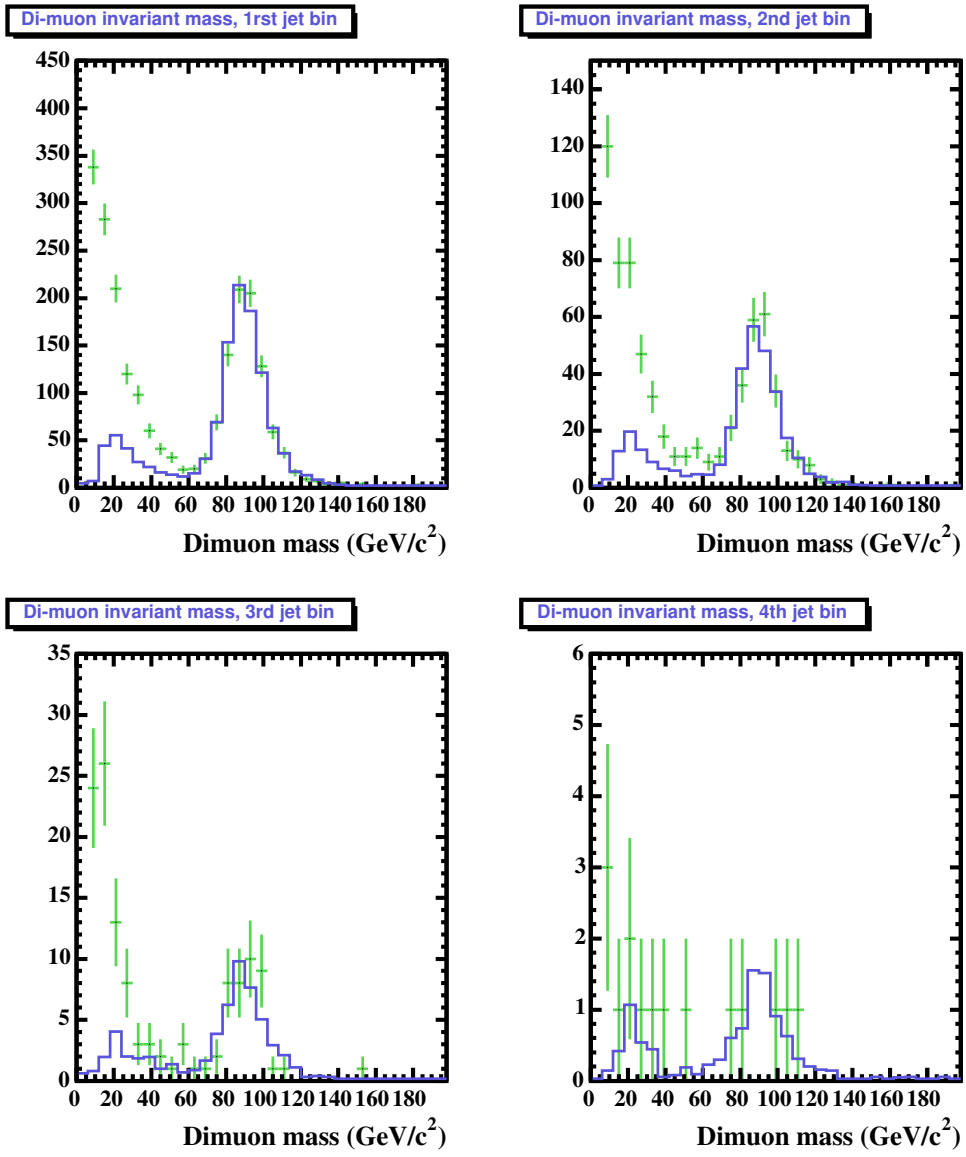


FIGURE 23 – Comparaison données/Monte Carlo des distributions en masse di-muons pour chaque multiplicité en jets (1, 2, 3 et 4 et plus). Les données sont représentées par les croix vertes et le Monte Carlo par les lignes continues bleues.

Les facteurs K_Z sont appliqués aux échantillons de Z se désintégrant en $\tau\tau$ et $\mu\mu$.

5.5.4 Bruit de fond multi-jets

Le bruit de fond multi-jets est évalué exclusivement à partir des données. Le nombre d'évènements de ce type sélectionnés par l'analyse est mesuré après étiquetage des jets de b

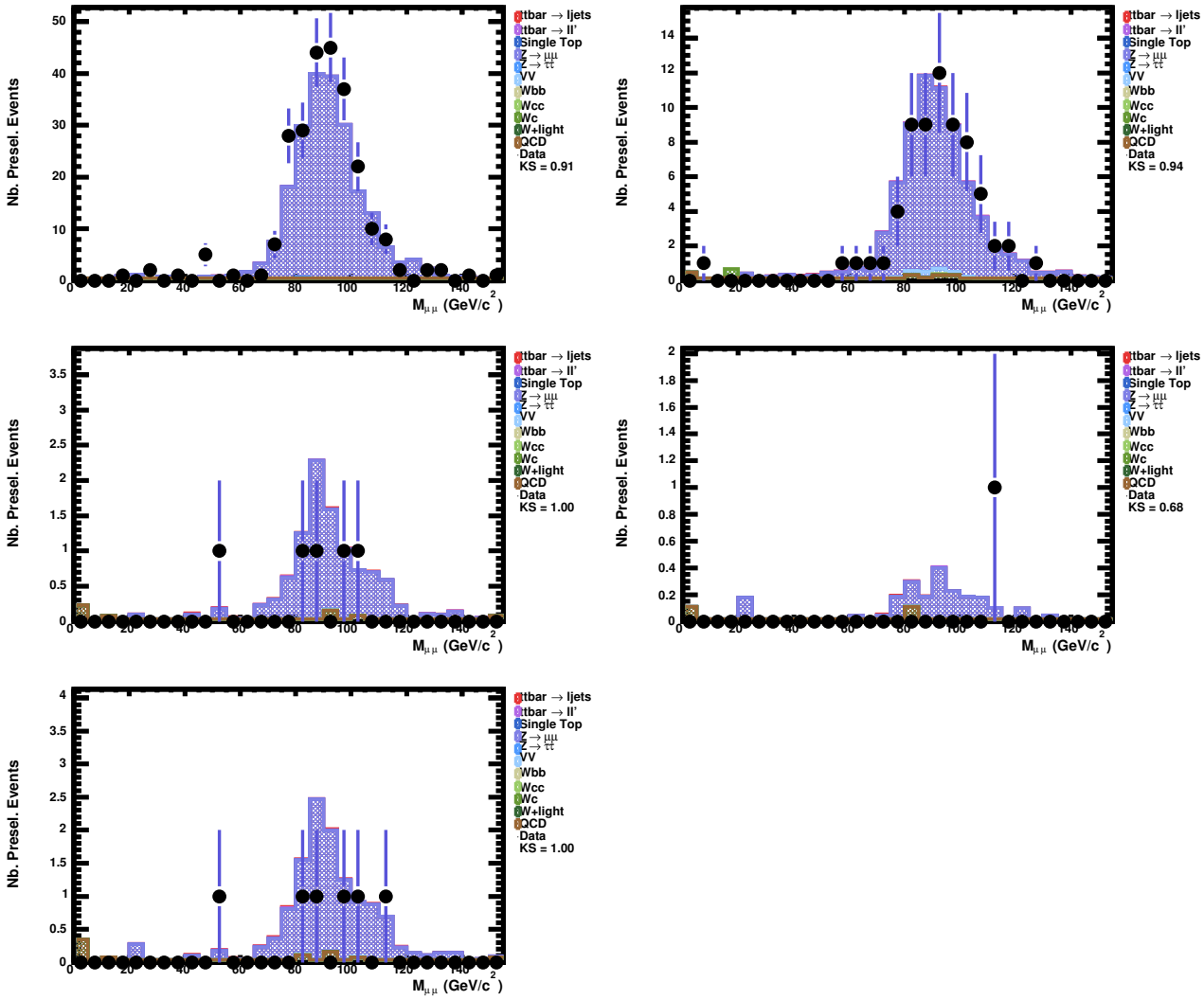


FIGURE 24 – Comparaison données/Monte Carlo des distributions en masse di-muons pour chaque multiplicité en jets (1, 2, 3, 4 et plus et 3 et plus) après la présélection stricte.

avec la méthode de la matrice pour chaque multiplicité en jet. Selon la méthode décrite précédemment, les évènements multi-jets sont considérés comme du bruit de fond, les évènements contenant un vrai lepton (top, Z et W) sont considérés comme du signal. La coupure stricte utilisée alors est celle définie dans le paragraphe détaillant la sélection : coupures sur l'isolation du muon pour le canal μ+jets et sur le maximum de vraisemblance permettant l'identification des électrons pour le canal e+jets. Le nombre d'évènements multi-jets est donc donné par :

$$\mathrm{N}_{\mathrm{multijets}}^{\mathrm{stricte}} = \varepsilon_{\mathrm{multijets}} \, \frac{\varepsilon_{\mathrm{signal}} \, \mathrm{N}_{\mathrm{lâche}} - \mathrm{N}_{\mathrm{stricte}}}{\varepsilon_{\mathrm{signal}} - \varepsilon_{\mathrm{multijets}}} \tag{5}$$

Les probabilités $\varepsilon_{\mathrm{signal}}$ et $\varepsilon_{\mathrm{multijets}}$ sont mesurées pour le canal électron et muon indépendamment après les coupures de présélection.

Dans les deux cas, $\varepsilon_{\mathrm{signal}}$ est mesuré à partir d'échantillons Monte Carlo W+jets et $\mathrm{t\bar{t}}$, et corrigé pour d'éventuelles différences avec les données. $\varepsilon_{\mathrm{multijets}}$ est mesuré directement sur les données dans une région à basse énergie transverse manquante ($\not\!\!\mathrm{E}_{\mathrm{T}} < 10$ GeV), là où le bruit de fond multi-jets domine.

Les critères de qualité sur l'identification des électrons dans le logiciel du système de déclenchement ayant évolué avec le temps, la mesure de $\varepsilon_{\mathrm{multijets}}$ dans le canal e+jets varie en fonction de ces versions. Aucune évolution significative n'est observée pour $\varepsilon_{\mathrm{signal}}$ en fonction des versions de ce logiciel. D'autre part, dans les deux canaux, on prend aussi en compte la légère dépendance de $\varepsilon_{\mathrm{signal}}$ et $\varepsilon_{\mathrm{multijets}}$ avec la multiplicité en jets des évènements. Dans le canal e+jets la variation de $\varepsilon_{\mathrm{multijets}}$ avec le nombre de jets est cependant négligeable. Les valeurs de ces différentes variables sont résumées dans la table 3.

TABLE 3 – Efficacité de la coupure stricte en % (coupure d'isolation pour le canal muon et sur l'identification de l'électron pour le canal électron) pour les échantillons W+jets et $\mathrm{t\bar{t}}$ (signal) et les données à basse $\not\!\!\mathrm{E}_{\mathrm{T}}$ (multijets). Ces efficacités sont données en fonction du nombre de jets dans l'évènement ou de la version du logiciel du système de déclenchement. Les incertitudes données sont uniquement statistiques.

	Canal μ+jets		Canal e+jets		
Nombre de jets	$\varepsilon_{\mathrm{signal}}$	$\varepsilon_{\mathrm{multijets}}$	$\varepsilon_{\mathrm{signal}}$	Version	$\varepsilon_{\mathrm{multijets}}$
$\mathrm{N}_{\mathrm{jet}} = 1$	89.72 ± 0.59	14.41 ± 0.30	$84,7 \pm 0,5$	v8-v11	$12,6 \pm 0,8$
$\mathrm{N}_{\mathrm{jet}} = 2$	85.38 ± 0.95	16.06 ± 0.75	$84,7 \pm 0,5$	v12	$18,6 \pm 0,9$
$\mathrm{N}_{\mathrm{jet}} = 3$	85.46 ± 0.70	17.76 ± 2.00	$83,0 \pm 0,6$	v13	$13,4 \pm 1,6$
$\mathrm{N}_{\mathrm{jet}} \geq 4$	81.77 ± 0.66	17.76 ± 2.00	$82,0 \pm 0,7$		

Échantillon multi-jets :
Afin de comparer les différentes distributions entre données et simulation, un échantillon d'évènement multi-jets est nécessaire. Cet échantillon est obtenu en appliquant aux données l'ensemble des coupures de sélection et en inversant celles correspondant à la coupure stricte.

5.5.5 W+jets

Le bruit de fond W+jets est évalué en plusieurs étapes : le nombre d'évènements W+jets est normalisé sur les données au niveau de la présélection, puis la probabilité d'étiquetage est mesurée avec des évènements simulés en fonction de la composition en saveur des jets qui accompagnent le W. La normalisation est donc obtenue en appliquant à nouveau la méthode de la matrice mais cette fois avant l'étiquetage des b. La méthode donne alors le nombre d'évènements contenant un vrai lepton :

$$\mathrm{N}_{\mathrm{signal}}^{\mathrm{stricte}} = \varepsilon_{\mathrm{signal}} \frac{\mathrm{N}_{\mathrm{stricte}} - \varepsilon_{\mathrm{multijets}}\, \mathrm{N}_{\mathrm{N_{lâche}}}}{\varepsilon_{\mathrm{signal}} - \varepsilon_{\mathrm{multijets}}} \tag{6}$$

À ces évènements, il faut soustraire les évènements dibosons, Z et top évalués grâce à la simulation pour obtenir une évaluation du nombre d'évènements W+jets avant l'étiquetage des b. On notera ici que la section efficace de production $\mathrm{t\bar{t}}$ que l'on cherche à obtenir entre en compte dans cette opération. On peut alors faire la mesure par itération en choisissant une première valeur par défaut (7 pb par exemple), mesurer la section efficace et itérer. La mesure converge rapidement : une seule itération est nécessaire pour obtenir un résultat stable. Néanmoins, une méthode différente sera utilisée pour la mesure finale, elle est décrite plus loin.

Échantillon W+jets :

Les évènements W+jets ont été simulés avec ALPGEN en différents échantillons en fonction de leur multiplicité en jets et de la saveur de ces derniers. Les différents échantillons et leur section efficace au premier ordre sont résumés dans le tableau 4.

TABLE 4 – Processus W+jets générés avec ALPGEN et leur section efficace. j correspond à un jet léger (u, d, s ou g) et J correspond à un jet qui n'est pas un jet b (soit un jet léger ou un jet c).

process	σ (pb)	process	σ (pb)	process	σ (pb)	process	σ (pb)	process	σ (pb)
Wj	1600	Wjj	517	Wjjj	163	Wjjjj	49.5		
Wc	51.8	Wcj	28.6	Wcjj	19.4	Wcjjj	3.15		
		$\mathrm{Wb\bar{b}}$	9.85	$\mathrm{Wb\bar{b}J}$	5.24	$\mathrm{Wb\bar{b}Jj}$	2.36	$\mathrm{Wb\bar{b}Jjj}$	0.939
		$\mathrm{Wc\bar{c}}$	24.3	$\mathrm{Wc\bar{c}J}$	12.5	$\mathrm{Wc\bar{c}Jj}$	5.83	$\mathrm{Wc\bar{c}Jjj}$	2.36

Pour éviter des doubles comptages, les évènements W+jets simulés sont combinés par multiplicité en jets en associant les partons aux jets reconstruits et en respectant les conditions suivantes (similaires à celles implémentées dans la méthode MLM [55]) :

1. La saveur d'un jet est déterminée par son association à un hadron dans un cône de rayon 0,5 : s'il y a au moins un hadron B dans le cône, le jet est identifié comme provenant d'un b, dans le cas contraire s'il y a un hadron charmé il est assimilé à un quark c et sinon il est défini comme un quark léger.
2. Dans le cas où il n'y a pas de désintégration de gluon dans l'élément de matrice du processus considéré (soit dans tous les cas sauf $\mathrm{W}b\bar{b} + \mathrm{X}$ et $\mathrm{W}c\bar{c} + \mathrm{X}$), on demande que le nombre de jets reconstruits soit égal au nombre de partons de l'élément de matrice du processus.

3. S'il y a désintégration d'un gluon (i.e. $\mathrm{W}b\bar{b} + \mathrm{X}$ et $\mathrm{W}c\bar{c} + \mathrm{X}$), le nombre de jets doit être égal au nombre N de parton ou à $\mathrm{N} - 1$; les partons issus de la désintégration d'un gluon peuvent donc être reconstruits en 1 ou 2 jets.
4. Le nombre de jets légers doit être égal au nombre de partons légers, sauf dans le cas où la multiplicité en jet considérée est supérieure à 4 (multiplicité inclusive).
5. On ne demande pas d'association des jets lourds aux partons sauf dans le cas de la production $\mathrm{Wc} + \mathrm{X}$.

Le résultat de cette classification est montré dans le tableau 5.

TABLE 5 – Classification des configurations en saveur des évènements W+jets en fonction de la multiplicité en jets. j correspond à l'un des partons $\mathrm{u}, \mathrm{d}, \mathrm{s}, \mathrm{g}$ et J à $\mathrm{u}, \mathrm{d}, \mathrm{s}, \mathrm{g}$ ou c. ($\mathrm{b\bar{b}}$) et ($\mathrm{\bar{c}}$) correspondent à des paires de quarks lourds reconstruits en un seul jet.

	W+1jet	W+2jets	W+3jets	W+≥4jets
W+jets légers	Wj	Wjj	Wjjj	Wjjjj
W+jets lourds		$\mathrm{Wb\bar{b}}$	$\mathrm{Wb\bar{b}J}$	$\mathrm{Wb\bar{b}Jj}$
		$\mathrm{Wc\bar{c}}$	$\mathrm{c\bar{c}J}$	$\mathrm{Wc\bar{c}Jj}$
	$\mathrm{W(b\bar{b})}$	$\mathrm{W(b\bar{b})j}$	$\mathrm{W(b\bar{b})jj}$	$\mathrm{W(b\bar{b})jjj}$
	$\mathrm{W(c\bar{c})}$	$\mathrm{W(c\bar{c})j}$	$\mathrm{W(c\bar{c})jj}$	$\mathrm{W(c\bar{c})jjj}$
	Wc	Wcj	Wcjj	Wcjjj

On peut ensuite mesurer la répartition de chaque combinaison de saveurs des quarks associés au W en fonction de la multiplicité en jets reconstruits, ce après la présélection. Les résultats sont regroupés dans le tableau 6. Ils permettent de mesurer les probabilités globales d'étiquetage en fonction de la multiplicité en jets, comme décrit dans le paragraphe traitant ce sujet.

TABLE 6 – Fraction de chaque échantillon W+jets de saveur de quarks donnée en fonction de la multiplicité en jets reconstruits. Mesure effectuée après la présélection et l'association jet-parton. Les incertitudes correspondent aux incertitudes statistiques liées aux nombres d'évènements Monte carlo produits.

Contribution	W+1jet	W+2jets	W+3jets	W+≥4jets
$\mathrm{Wb\bar{b}}$		$(1.23 \pm 0,08)\%$	$(2,05 \pm 0,21)\%$	$(2,84 \pm 0,16)\%$
$\mathrm{Wc\bar{c}}$		$(1,69 \pm 0,12)\%$	$(2,94 \pm 0,37)\%$	$(4,44 \pm 0,29)\%$
$\mathrm{W(b\bar{b})}$	$(0,86 \pm 0,03)\%$	$(1,46 \pm 0,09)\%$	$(2,03 \pm 0,15)\%$	$(2,99 \pm 0,24)\%$
$\mathrm{W(c\bar{c})}$	$(1,23 \pm 0,05)\%$	$(2,26 \pm 0,15)\%$	$(3,08 \pm 0,24)\%$	$(5,06 \pm 0,54)\%$
Wc	$(4,41 \pm 0,18)\%$	$(6,25 \pm 0,43)\%$	$(4,93 \pm 0,48)\%$	$(4,30 \pm 0,23)\%$
W + jets	$(93,50 \pm 0,20)\%$	$(87,10 \pm 0,70)\%$	$(84,96 \pm 1,12)\%$	$(80,36 \pm 0,64)\%$

5.5.6 Résultats

Afin d'obtenir le nombre d'évènements multi-jets et W+jets, la méthode de la matrice est donc appliquée deux fois : une fois au niveau de la présélection stricte et la deuxième après l'étiquetage des jets. Le nombre d'évènements avec un vrai lepton au niveau de la présélection représente le nombre d'évènements de type W associé à des jets. Pour obtenir le nombre d'évènements de bruit de fond W+jets, il faudra retrancher à ce nombre le nombre d'évènements $\mathrm{t\bar{t}}$. Le nombre d'évènements avec un faux lepton après étiquetage représente lui directement notre évaluation du nombre d'évènement multi-jets après l'étiquetage. La table 7 résume donc les nombres d'évènements sélectionnés dans les données au niveau de la présélection et de l'étiquetage des b ainsi que l'évaluation de leur répartition en évènements contenant un vrai ou un faux lepton.

TABLE 7 – Nombre d'évènements sélectionnés dans les données $\mathrm{N_{données}}$ et résultats de la méthode de la matrice évaluant la répartition des évènements contenant un vrai, $\mathrm{N_{vrai\ lepton}}$, ou un faux lepton, $\mathrm{N_{faux\ lepton}}$. Chacun de ces nombres est donné au niveau de la présélection stricte et de l'étiquetage des jets de b, pour les canaux e+jets et μ+jets et en fonction de la multiplicité en jets.

Sample	e+jets				μ+jets			
	1 jet	2 jets	3 jets	≥ 4 jets	1 jet	2 jets	3 jets	≥ 4 jets
Résultats après la présélection stricte								
$\mathrm{N_{données}}$	6153	2217	466	119	6827	2267	439	100
$\mathrm{N_{vrai\ lepton}}$	5805.6±83.2	1975.8±50.0	394.9±23.0	99.8±11.6	6607.0±84.9	2155.1±49.9	405.5±22.1	91.4±10.7
$\mathrm{N_{faux\ lepton}}$	347.4±17.7	241.2±11.0	71.1±4.7	19.2±2.3	220.0±10.2	111.9±8.2	33.5±4.2	8.6±1.9
Résultats après l'étiquetage des jets								
$\mathrm{N_{données}}$	59	38	17	17	64	45	21	7
$\mathrm{N_{vrai\ lepton}}$	55.5±8.0	35.1±6.5	14.5±4.4	17.1±4.3	60.8±8.2	43.8±7.0	20.0±4.8	4.9±2.9
$\mathrm{N_{faux\ lepton}}$	3.5±1.1	2.9±0.9	2.5±0.8	< 0.01	3.2±0.9	1.2±0.8	1.0±0.7	2.1±0.8

5.6 Mesure de la section efficace

La section efficace est obtenue à partir des mesures des nombres d'évènements de signal et de bruits de fond décrits précédemment. Le principe de la mesure est décrit dans le paragraphe suivant, les détails de l'optimisation faite avec une méthode de maximum de vraisemblance dans un second temps.

5.6.1 Principe

Après avoir appliqué la sélection et les coupures liées à l'étiquetage des b aux échantillons de données dans les canaux électron et muon avec 3 et 4 ou plus jets associés, et évalué les bruits de fond résiduels, la section efficace $\mathrm{t\bar{t}}$ est obtenue de la façon suivante dans chacun des 4 canaux :

$$\sigma(\mathrm{t\bar{t}}) = \frac{\mathrm{N_{t\bar{t}}}}{\varepsilon \times \mathrm{Br} \times \mathcal{L}\mathrm{umi}} = \frac{\mathrm{N_{données}} - \mathrm{N_{bruit\ de\ fond}}}{\varepsilon \times \mathrm{Br} \times \mathcal{L}\mathrm{umi}} \quad (7)$$

$\mathrm{N_{données}}$ correspond au nombre d'évènements qui passent la sélection stricte et l'étiquetage des b, on le note $\mathrm{N^{tag}_{stricte}}$. $\mathrm{N_{bruit\ de\ fond}}$ est la somme de tous les évènements de bruit de

fond qui passe cette même sélection. Ce nombre est fonction du nombre d'évènements de données qui passent les sélections lâche et stricte, et qui sont étiquetés ou non ($\mathrm{N}^{\mathrm{sel}}_{\mathrm{stricte}}$, $\mathrm{N}^{\mathrm{sel}}_{\mathrm{N_{lâche}}}$, $\mathrm{N}^{\mathrm{tag}}_{\mathrm{stricte}}$ and $\mathrm{N}^{\mathrm{tag}}_{\mathrm{N_{lâche}}}$)[9] via la mesure des bruits de fond multi-jets et W+jets avec la méthode de la matrice . On ajoute un indice j à ces 4 variables qui correspond à l'un des 4 canaux considérés : c'est-à-dire le canal électron et le canal muon, chacun selon une multiplicité de 3 jets ou de 4 jets ou plus. $\mathrm{N}_{\text{bruit de fond}}$ dépend aussi de la section efficace $\mathrm{t\bar{t}}$ ($\sigma(\mathrm{t\bar{t}})$) via l'évaluation du bruit de fond W+jets.

Les estimateurs des vraies valeurs de $\mathrm{N_j}$ et de $\sigma(\mathrm{t\bar{t}})$ sont obtenues à partir des valeurs observées et mesurées des $\mathrm{N_j}$. Ces estimateurs, notés $\tilde{\mathrm{N}}_{\mathrm{j}}$, sont obtenus en maximisant la fonction de vraisemblance $\mathcal{L}_1(\mathrm{N}, \sigma(\mathrm{t\bar{t}}))$ en fonction de $\mathrm{N_j}$ et $\mathrm{sigma}(\mathrm{t\bar{t}})$:

$$\mathcal{L}_1(\mathrm{N_j}, \sigma(\mathrm{t\bar{t}})) = \prod_{\mathrm{j}=1}^{4} \mathrm{F_j}(\tilde{\mathrm{N}}_{\mathrm{j}} | \mathrm{N_j}, \sigma(\mathrm{t\bar{t}})) \tag{8}$$

$\mathrm{F_j}$ est la densité de probabilité de $\mathrm{N_j}$.

Les valeurs des variables $\mathrm{N_j}$ et $\sigma(\mathrm{t\bar{t}})$ sont laissées libres pendant le processus de maximisation mais les $\mathrm{N_j}$ sont contraintes à leur valeur mesurée respective $\tilde{\mathrm{N}}_{\mathrm{j}}$ via la densité de probabilité $\mathrm{F_j}$. Notre estimation de la section efficace $\mathrm{t\bar{t}}$ sera la valeur de $\sigma(\mathrm{t\bar{t}})$ qui maximise la fonction de vraisemblance.

5.6.2 Construction du maximum de vraisemblance

La fonction de vraisemblance telle que construite ci-dessus n'est pas satisfaisante car les variables $\mathrm{N_j}$ ne représentent pas des échantillons indépendants (disjoints). Par exemple les évènements issus de la sélection stricte sont inclus dans ceux issus de la sélection lâche. Nous allons donc répartir les évènements dans des échantillons disjoints contenant seulement ceux ayant passés la coupure stricte $\mathrm{M_{stricte}}$ ou ceux ayant passés la coupure lâche mais pas la coupure stricte $\mathrm{M_{lâche-stricte}}$. Ces échantillons seront aussi divisés en 2 ensembles disjoints : ceux avec au moins un jet étiqueté b ($\mathrm{M}^{\geq 1}_{\mathrm{stricte}}$ et $\mathrm{M}^{\geq 1}_{\mathrm{lâche-stricte}}$) et ceux sans aucun jet étiqueté ($\mathrm{M}^{0}_{\mathrm{stricte}}$ et $\mathrm{M}^{0}_{\mathrm{lâche-stricte}}$).

Ces nouvelles variables s'expriment facilement en fonction des variables mesurées par l'analyse :

$$\begin{aligned}
\mathrm{M}^{0}_{\mathrm{stricte}} &= \mathrm{N}^{\mathrm{sel}}_{\mathrm{stricte}} - \mathrm{N}^{\mathrm{tag}}_{\mathrm{stricte}} \\
\mathrm{M}^{\geq 1}_{\mathrm{stricte}} &= \mathrm{N}^{\mathrm{tag}}_{\mathrm{stricte}} \\
\mathrm{M}^{0}_{\mathrm{lâche-stricte}} &= (\mathrm{N}^{\mathrm{sel}}_{\mathrm{lâche}} - \mathrm{N}^{\mathrm{sel}}_{\mathrm{stricte}}) - (\mathrm{N}^{\mathrm{tag}}_{\mathrm{lâche}} - \mathrm{N}^{\mathrm{tag}}_{\mathrm{stricte}}) \\
\mathrm{M}^{\geq 1}_{\mathrm{lâche-stricte}} &= \mathrm{N}^{\mathrm{tag}}_{\mathrm{lâche}} - \mathrm{N}^{\mathrm{tag}}_{\mathrm{stricte}}
\end{aligned} \tag{9}$$

Ces 4 nouvelles variables seront notées en fonction du canal j considéré : $\mathrm{M_j}$ si l'on considère leur vraie valeur et $\tilde{\mathrm{M}}_{\mathrm{j}}$ pour leur valeur mesurée. On obtient alors la fonction de vraisemblance $\mathcal{L}_1$ suivante :

$$\mathcal{L}_1(\mathrm{M_j}, \sigma(\mathrm{t\bar{t}})) = \prod_{\mathrm{j}=1}^{4} \mathrm{G_j}(\tilde{\mathrm{M}}_{\mathrm{j}} | \mathrm{M_j}, \sigma(\mathrm{t\bar{t}})) \tag{10}$$

9. avec l'indice *sel* pour les évènements qui passent la présélection, l'indice *tag* pour les évènements étiquetés et les indices *stricte* ou *lâche* pour préciser le type de présélection.

où les G_j correspondent aux densités de probabilité des M_j. Les paramètres de la fonction de vraisemblance ainsi définie qui seront ajustés dans la procédure de maximisation sont donc les 4 variables M_j et $\sigma(t\bar{t})$.

Les variables M_j seront contraintes vis à vis de leurs valeurs mesurées $\tilde{M}_j$ dans la procédure de maximisation, en considérant qu'elles suivent une distribution de Poisson :

$$\begin{aligned} G_j(\tilde{M}_j/M_j, \sigma(t\bar{t})) &= \mathcal{P}(\tilde{M}^0_{j,\text{ stricte}}, M^0_{j,\text{ stricte}}) \times \mathcal{P}(\tilde{M}^{\geq 1}_{j,\text{ stricte}}, M^{\geq 1}_{j,\text{ stricte}}) \\ &\quad \times \mathcal{P}(\tilde{M}^0_{j,\ M_{\text{lâche}}-\text{stricte}}, M^0_{j,\text{ stricte}}) \\ &\quad \times \mathcal{P}(\tilde{M}^{\geq 1}_{j,\ M_{\text{lâche}}-\text{stricte}}, M^{\geq 1}_{j,\text{ stricte}}) \end{aligned} \tag{11}$$

où $\mathcal{P}(\tilde{M}, M) = \frac{e^{-M} \cdot M^{\tilde{M}}}{\tilde{M}!}$ est la fonction de distribution de Poisson.

On notera que $\sigma(t\bar{t})$ n'apparaît pas explicitement dans l'équation ci-dessus mais elle entre en compte dans l'expression de M_j car $\sigma(t\bar{t})$ dépend directement de $N^{tag}_{j,t\bar{t},\text{stricte}}$ et les M_j correspondant aux coupures strictes sont données par :

$$M^{\geq 1}_{j,\text{ stricte}} = N^{tag}_{j,t\bar{t},\text{ stricte}} + N^{tag}_{j,\text{bruit de fond, stricte}} \tag{12}$$

qui dit simplement que le nombre total d'évènements ayant passé la présélection stricte et ayant été étiquetés est égale à la somme des évènements $t\bar{t}$ et de bruit de fond ayant passés cette même sélection.

D'autre part, il faut noter qu'il existe une relation qui lie les 4 variables M_j et $\sigma(t\bar{t})$ via les équations de la méthode de la matrice. En développant l'équation 12, on obtient en effet (voir annexe A pour les détails) :

$$\begin{aligned} N^{\geq 1}_{\text{stricte}} &= a\, N^0_{\text{stricte}} - ab\, N^0_{N_{\text{lâche}}\ -\ \text{stricte}} + b\, N^{\geq 1}_{N_{\text{lâche}}\ -\ \text{stricte}} \\ &\quad + c\,(P^{tag}_{\text{MCbdf}} - P^{tag}_{W+\text{jets}})\, N^{sel}_{\text{MCbdf, stricte}} \\ &\quad + c\,(P^{tag}_{t\bar{t}} - P^{tag}_{W+\text{jets}})\, N^{sel}_{t\bar{t},\text{ stricte}}(\sigma(t\bar{t})) \end{aligned} \tag{13}$$

où

$$\begin{aligned} a &= \frac{P^{tag}_{W+\text{jets}}}{1 - P^{tag}_{W+\text{jets}}} \\ b &= \frac{\varepsilon_{\text{multijets}}}{1 - \varepsilon_{\text{multijets}}} \\ c &= \frac{\varepsilon_{\text{sig}} - \varepsilon_{\text{multijets}}}{\varepsilon_{\text{sig}} \cdot (1 - \varepsilon_{\text{multijets}}) \cdot (1 - P^{tag}_{W+\text{jets}})} \end{aligned} \tag{14}$$

où P^{tag} représente la probabilité d'étiquetage des différents processus, et les ε les probabilités de passer les coupures strictes des différents processus de la méthode de la matrice.

Ainsi, comme il existe une relation entre les 5 paramètres de la fonction de vraisemblance (équation (22)), la fonction de vraisemblance ne doit être maximisée qu'en fonction

de 4 d'entre eux. On choisit de fixer $\mathrm{N}^{\geq 1}_{\mathrm{stricte}}$ selon l'équation (22).

En résumé, on obtient une fonction de vraisemblance $\mathcal{L}_1(\mathrm{M_j}, \sigma(\mathrm{t\bar{t}}))$ construite à partir des équations (10) et (11) qui dépend de 4 paramètres indépendants qui seront laissés libres dans la procédure de maximisation. L'un de ces paramètres libres est la section efficace $\mathrm{t\bar{t}}$ que l'on veut mesurer. Notre mesure de section efficace correspondra donc à celle qui maximise la fonction de vraisemblance ainsi définie.

5.7 Les erreurs systématiques

5.7.1 Prise en compte des erreurs systématiques

Les erreurs systématiques relatives à l'analyse sont prises en compte dans la mesure de la section efficace au moment du calcul du maximum de vraisemblance. En effet, chaque erreur systématique est introduite dans la fonction de vraisemblance comme un nouveau paramètre appelé paramètre de nuisance. Ces paramètres sont laissés libres pendant la procédure de maximisation mais sont contraints de suivre une distribution gaussienne centrée en 0 et d'écart-type égal à l'erreur systématique mesurée. Chaque efficacité de sélection ou probabilité d'étiquetage est alors recalculée en fonction des valeurs de chaque paramètre de nuisance associé à une systématique donnée. Lorsqu'un même paramètre affecte plusieurs variables de la fonction de vraisemblance, ces variables sont recalculées en conséquence, ce qui permet de prendre en compte les corrélations liées aux systématiques de façon naturelle.

Pour obtenir la section efficace, on multiplie donc la fonction de vraisemblance $\mathcal{L}_1$ décrite dans la section 5.6.2 par la fonction $\mathcal{L}_2$ suivante

$$\mathcal{L}_2 = \prod_{\mathrm{j}} \mathcal{G}(\nu_{\mathrm{j}}; 0, \sigma_{\mathrm{j}}) \tag{15}$$

où les $\mathcal{G}(\nu_{\mathrm{j}}; 0, \sigma_{\mathrm{j}})$ sont les probabilités gaussiennes que le paramètre de nuisance j prenne la valeur ν_{j}.

La maximisation du produit des deux fonctions de vraisemblance se fait alors sur environ 200 paramètres. Le résultat donne une mesure de la section efficace à laquelle est associée une erreur qui inclue à la fois les erreurs statistiques et systématiques.

Cependant pour avoir une idée de l'impact de chaque systématique (ou de chaque groupe de systématiques) séparément, on peut figer tous les paramètres de nuisance sauf un.

5.7.2 Évaluation des erreurs systématiques

Incertitude sur la luminosité

La luminosité correspondant aux données enregistrées est mesurée par la collaboration DØ à partir de la mesure de la section efficace inélastique $p\overline{p}$. L'erreur relative associée est de 6.1 % [56].

Incertitudes liées à la présélection

Lorsque l'efficacité de sélection du signal et des bruits de fond est mesurée à partir de la simulation, les incertitudes correspondantes sont liées d'une part à la statistique des échantillons utilisés et d'autre part aux différences de comportement de la simulation par rapport

à la réalité. Ces incertitudes sont évaluées pour chaque échantillon et chaque variable utilisée. Le traitement des jets fait l'objet d'une procédure particulière [57].

Incertitudes liées à l'étiquetage

L'évaluation de cette incertitude est décomposée en plusieurs étapes :
1 - Identification des muons dans les jets :
L'efficacité de sélection d'un tel muon est mesurée dans les données sur un échantillon $\mathrm{J}/\Psi \to \mu\mu$ [53]. Aucune différence n'a été relevée entre données et simulation en fonction des variables η, ϕ ou $\mathrm{p_T}$. L'erreur prend donc en compte les erreurs statistiques afférentes ainsi que les erreurs liées à la procédure d'identification des muons.
2- Taux d'étiquetage des jets issus de partons légers :
Il s'agit de l'incertitude systématique dominante de l'analyse. Le taux d'étiquetage est mesuré en 2 étapes [53] :

- La première consiste à mesurer l'efficacité d'étiquetage avec les données sur un échantillon multi-jets avec au moins un objet électromagnétique ;
- La deuxième doit soustraire les composants en saveur lourde de cet échantillon.

Le taux d'étiquetage des jets légers est donc calculé de la façon suivante :

$$\epsilon^{\text{léger}}_{\text{données}} = \epsilon_{\text{données}} - \mathrm{Fraction(b)} \times \varepsilon_{\mathrm{b}} \times \mathrm{SF} - \mathrm{Fraction(c)} \times \varepsilon_{\mathrm{c}} \times \mathrm{SF}$$

où $\epsilon^{\text{léger}}_{\text{données}}$ est le taux d'étiquetage des jets légers que nous voulons mesurer, $\mathrm{epsilon_{données}}$ est l'efficacité d'étiquetage global de l'échantillon considéré. Fraction(b) et Fraction(c) sont les fractions inclusives de jet de b ou de c dans les évènements Monte Carlo di-jets. ϵ_{b} and ϵ_{c} sont les efficacités d'étiquetage de jets issus de quark b ou c et mesurées sur les échantillons $\mathrm{QCD} \to \mathrm{b\bar{b}}$ et $\mathrm{Z} \to \mathrm{b\bar{b}}$. SF est le facteur de correction entre données et simulation de l'efficacité d'identification des muons.

Les contributions des différents termes ci-dessus dans l'incertitude systématique mesurée sont les suivantes : différence des taux d'étiquetage données/simulation (9,0 %), erreur statistique (2,0 %), approximation liée à la fraction de saveur prise comme constante (6,7 % et 10,6 % respectivement pour les fractions de b et de c), différence d'efficacité d'étiquetage entre les échantillons Monte Carlo $\mathrm{QCD} \to \mathrm{b\bar{b}}$ et $\mathrm{Z} \to \mathrm{b\bar{b}}$ (44,4 %). Les incertitudes sur le mauvais taux d'étiquetage dans les données et le Monte Carlo sont respectivement 34,5 % et 5,5 %. Elles correspondent à une incertitude systématique de 35 % sur le facteur de correction.

Incertitude liée à l'évaluation des bruits de fond avec les données

Les incertitudes liées à la méthode de la matrice et à la normalisation du bruit de fond Z sont obtenues en faisant varier les paramètres utilisés (ϵ_{sig}, ϵ_{QCD}, $\mathrm{K_Z}$) par une déviation standard. Pour le bruit de fond Z+jets, on tient compte aussi de la statistique limitée (données et simulation) utilisée pour construire la distribution de la masse invariante di-muons.

En ce qui concerne le bruit de fond W+jets, l'évaluation est un peu plus complexe : il est nécessaire en effet de prendre en compte les incertitudes liées aux fractions en multiplicité des jets associés au W en fonction de leur saveur. Ces fractions dépendent des section efficaces effectives W+jets obtenues avec ALPGEN et sont modifiées avec l'association des jets aux partons et corrigées des différences entre données et simulation en ce qui concerne la

sélection des évènements. L'incertitude systématique sur le bruit de fond W+jets tient donc compte de l'ensemble de ces opérations :

- incertitudes théoriques sur les sections efficaces d'ALPGEN provenant des fonctions de densité partonique, des choix des échelles de factorisation et de renormalisation et de l'incertitude sur les masses des quarks lourds ainsi que de l'incertitude des facteurs de normalisation à l'ordre NLO ;
- incertitude sur la méthode d'association des jets aux partons ; cette incertitude est évalué en comparant la méthode utilisée avec la méthode MLM ;
- incertitude sur la sélection pour laquelle on procède comme pour les autres bruits de fond.

6 Résultats

En considérant séparément les canaux électron et muon, on obtient les sections efficaces suivantes :

$$\begin{aligned} \mathrm{e+jets} \ : \ \sigma_{\mathrm{p\bar{p}}\to\mathrm{t\bar{t}+X}} &= 8.7\ ^{+2.5}_{-2.2}\ (\mathrm{stat+syst})\ \pm 0.5\ (\mathrm{lumi})\ \mathrm{pb}, \\ \mu\mathrm{+jets} \ : \ \sigma_{\mathrm{p\bar{p}}\to\mathrm{t\bar{t}+X}} &= 4.2\ ^{+2.9}_{-2.6}\ (\mathrm{stat+syst})\ \pm 0.3\ (\mathrm{lumi})\ \mathrm{pb}. \end{aligned} \tag{16}$$

En utilisant l'ensemble des canaux, on obtient :

$$\sigma_{\mathrm{p\bar{p}}\to\mathrm{t\bar{t}+X}} = \ 7.3\ ^{+2.0}_{-1.8}\ (\mathrm{stat+syst})\ \pm 0.4\ (\mathrm{lumi})\ \mathrm{pb}.$$

La table 8 résume la composition des échantillons de données après l'étiquetage des évènements. La contribution des évènements $\mathrm{t\bar{t}}$ est calculée en utilisant la section efficace de production $\mathrm{t\bar{t}}$ telle qu'elle a été mesurée.

Les figures 25 et 26 montrent le nombre d'évènements étiquetés, observés et prédits, pour chaque multiplicité de jets. Les canaux e+jets et μ+jets sont représentés séparément sur la première figure et sont combinés sur la deuxième.

Les distributions cinématiques des évènements de données réelles sélectionnés et étiquetés sont comparées à la somme des bruits de fond évalués et du signal $\mathrm{t\bar{t}}$. Globalement et pour chaque canal, ces comparaisons montrent un accord raisonnable. Quelques unes de ces distributions sont montrées en annexe B.

Enfin, les contributions des principales sources d'incertitude systématique sur la mesure de la section efficace sont regroupées dans le tableau 9.

7 Mise en perspective

La mesure décrite ici est en accord avec les calculs théoriques et avec les deux autres mesures effectuées pour la même période de prise de données mais avec des méthodes différentes. L'analyse purement topologique [58] mesure en effet :

$$\sigma_{\mathrm{p\bar{p}}\to\mathrm{t\bar{t}+X}} = 6,4\ ^{+1,3}_{-1,2}\ (\mathrm{stat})\ \pm 0,7\ (\mathrm{syst})\ \pm 0,4\ (\mathrm{lumi})\ \mathrm{pb},$$

et l'analyse utilisant un étiquetage des b basé sur le temps de vie des hadrons B [59] :

TABLE 8 – Nombre d'évènements sélectionnés dans les données et évaluation des bruits de fond dans les deux canaux e+jets et μ+jets en fonction de la multiplicité en jets. Le nombre d'évènements $t\bar{t}$ est obtenu en utilisant la section efficace mesurée. Les incertitudes sont les incertitudes statistiques uniquement.

Sample	1 jet	2 jets	3 jets	≥ 4 jets
W+jets légers	62.0±9.8	17.2±6.1	5.5±1.3	0.63±0.29
Wc	23.3±2.3	10.3±1.7	1.49±0.33	0.23±0.06
$Wc\bar{c}$	6.30±0.7	8.3±1.0	1.1±0.2	0.46±0.13
$Wb\bar{b}$	9.7±0.5	11.7 ±0.7	3.1±0.3	0.45±0.08
W+jets	101.3±10.1	47.4±6.4	11.2±1.4	1.77±0.33
Multi-jets	6.7±1.4	4.10±1.22	3.50±1.07	2.01±0.85
tb	0.53±0.02	2.18±0.05	0.28±0.02	0.02±0.01
tqb	0.91±0.06	2.21±0.09	0.92±0.06	0.13±0.02
diboson	1.83±0.25	3.05±0.30	0.32±0.10	< 0.01
$Z \to \tau^+\tau^- + \text{jets}$	0.78±0.52	0.22±0.15	0.26±0.15	< 0.01
$Z \to \mu^+\mu^- + \text{jets}$	11.5±1.1	9.6±0.6	2.84±0.45	1.10±0.28
bruit de fond	123.4±10.2	68.8±6.6	19.3±1.8	4.0±1.0
$t\bar{t} \to l + \text{jets}$	0.49±0.05	6.6±0.2	20.4±0.4	16.6±0.3
$t\bar{t} \to ll$	1.18±0.04	6.0±0.1	2.16±0.06	0.31±0.02
total	125.1±10.2	81.4±6.6	41.9±1.8	22.0±1.0
données	123	83	38	24

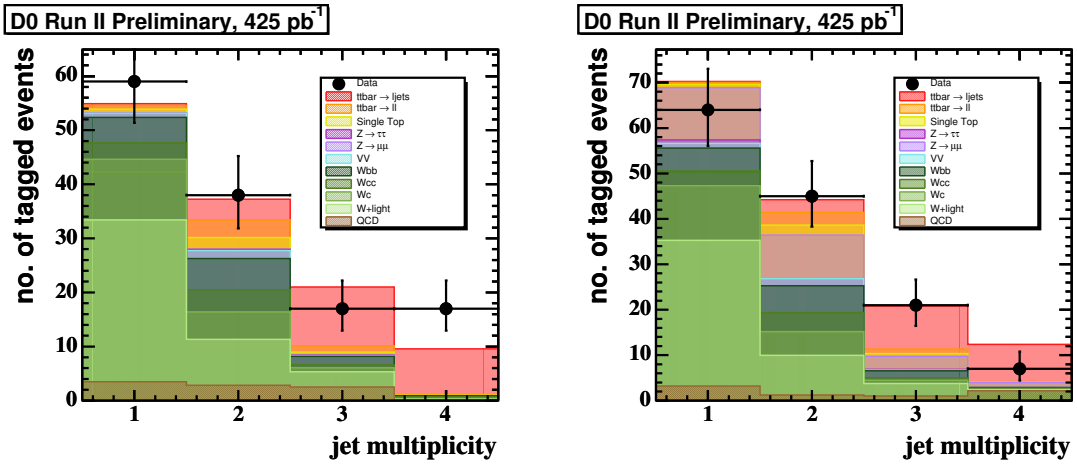


FIGURE 25 – Comparaison entre le nombre d'évènements observés et prédits et composition du bruit de fond en fonction de la multiplicité en jets pour le canal électron (à gauche) et le canal muon (à droite). La prédiction sur le signal $t\bar{t}$ est faite en utilisant la section efficace mesurée (sur les deux canaux combinés).

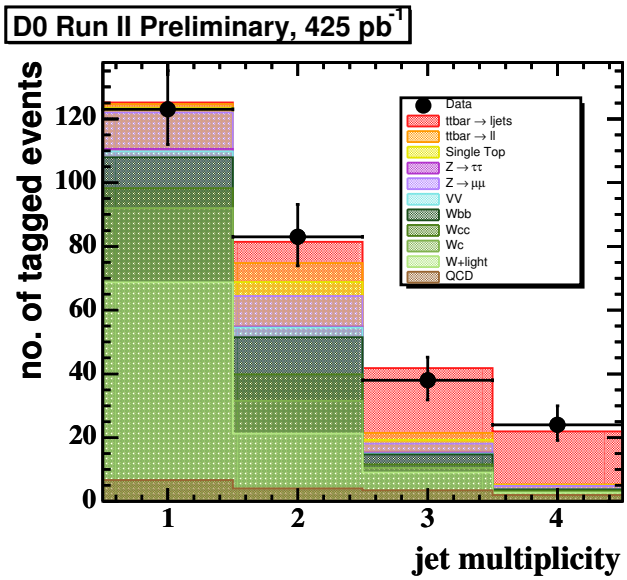


FIGURE 26 – Comparaison entre le nombre d'évènements observés et prédits et composition du bruit de fond en fonction de la multiplicité en jets pour les canaux e+jets et μ+jets combinés. La prédiction sur le signal $t\bar{t}$ est faite en utilisant la section efficace mesurée.

TABLE 9 – Impact des principales incertitudes systématiques sur la section efficace mesurée pour les deux canaux électron et muon combinés.

Source	σ_+(pb)	σ_-(pb)
Présélection μ+jets	+0.18	-0.13
Présélection e+jets	+0.19	-0.13
Déclenchement EM	+0.00	-0.03
Déclenchement Muon	+0.12	-0.09
Déclenchement Jet	+0.00	-0.04
Échelle d'énergie des jets	+0.19	-0.12
Résolution en énergie des jets	+0.02	-0.02
Identification des jets	+0.14	-0.12
Normalization Z+jets	+0.06	-0.07
Étiquetage saveurs lourdes	+0.24	-0.17
Taux de faux étiquetage	+0.84	-0.78
Méthode de la matrice	+0.33	-0.35
Statistique Monte Carlo	+0.25	-0.27
Fractions W	+0.13	-0.19
Systématique totale (somme quadratique)	+1.04	-0.98

$$\sigma_{p\bar{p}\to t\bar{t}+X} = 6,6 \pm 0,9\,(\text{stat}+\text{syst}) \pm 0.4\,(\text{lumi})\ \text{pb}.$$

L'ensemble de ces mesures ainsi que celles effectuées dans les autres états finaux sont regroupées sur la figure 27 à gauche. On observe un bon accord entre les mesures elles-mêmes et entre les mesures et les calculs théoriques. L'ensemble de ces résultats ont depuis été mis à jour avec les données que l'expérience a continué à accumuler. Ils sont présentés sur la même figure à droite. L'expérience utilise toujours deux méthodes : une méthode topologique et une méthode utilisant l'étiquetage des b. Cet étiquetage est maintenant basé sur un réseau de neurones qui prend en compte plusieurs propriétés des jets de b (vertex déplacé et paramètre d'impact des traces). Grâce à l'accumulation de la statistique et à des outils améliorés, les incertitudes sur la section efficace ont diminué significativement. Dans le canal lepton+jets, avec 5,4 fb^{-1}, la mesure combinée la plus récente est de :

$$\sigma_{p\bar{p}\to t\bar{t}+X} = 7,65\,^{+0,25}_{-0,25}\,(\text{stat})\,^{+0,75}_{-0,57}\,(\text{syst})\ \text{pb}.$$

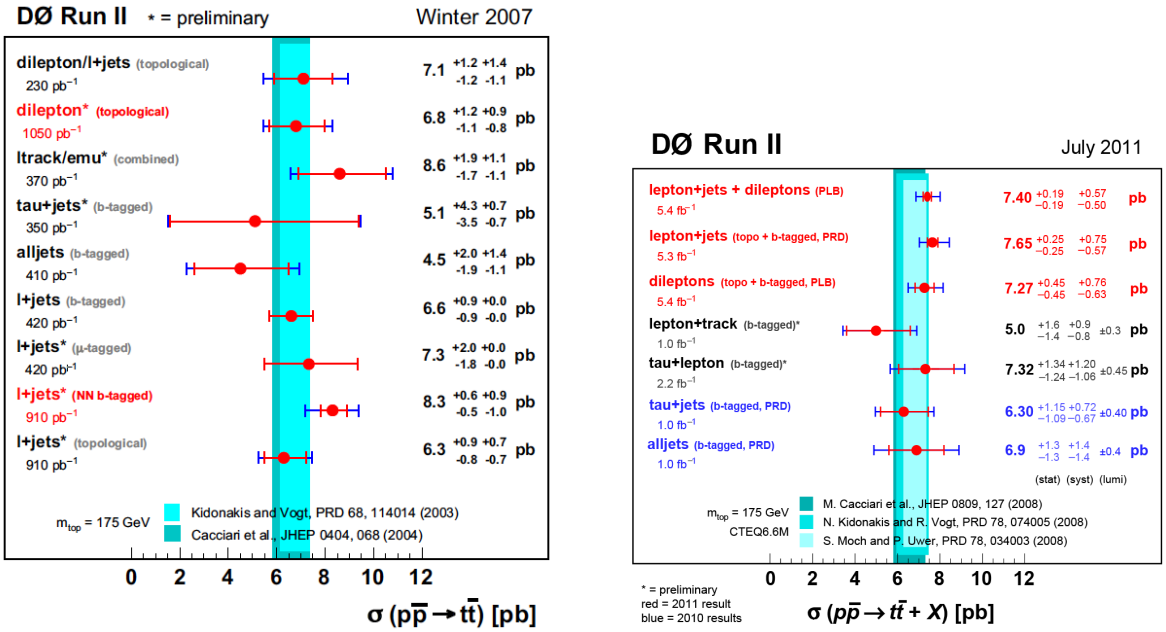


FIGURE 27 – Compilation des mesures par la collaboration DØ de la section efficace $t\bar{t}$ dans les différents canaux effectuées. Compilation de 2007 à gauche et de 2011 à droite.

Depuis le démarrage du LHC, ATLAS et CMS ont elles-aussi mesuré la section efficace de production $t\bar{t}$ mais pour des collisions proton-proton et à une énergie de 7 et 8 TeV. À ces énergies, c'est la production forte de top par fusion de gluons qui domine. Comme illustré sur la figure 28, les mesures au LHC comme celles effectuées au Tevatron sont en bon accord avec les calculs théoriques. Les mesures effectuées dans les différents états finaux sont elles aussi totalement compatibles.

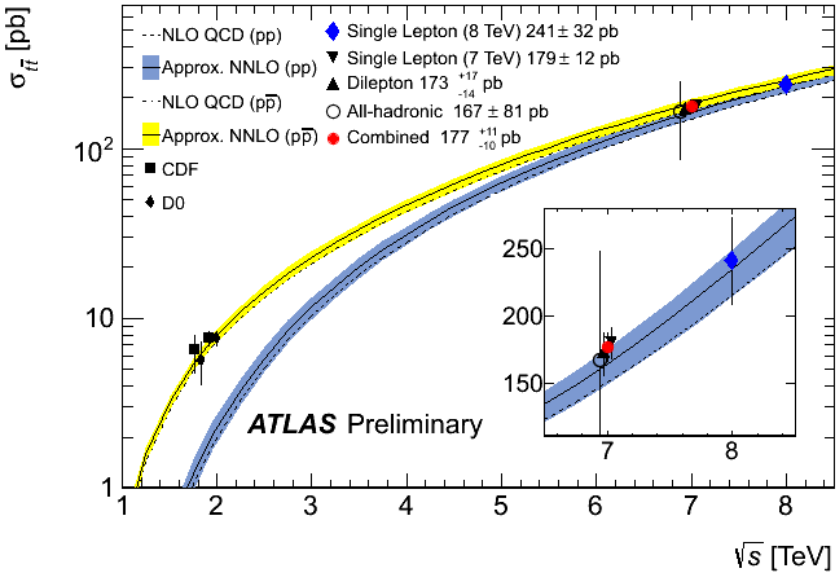


FIGURE 28 – Mesure de la section efficace de production $t\bar{t}$ au Tevatron et au LHC et comparaison aux calculs théoriques en fonction de l'énergie des collisions.

Quatrième partie
Projet de recherche

8 Introduction

L'ensemble des systèmes et applications liés au traitement des données du LHC et de la grille de calcul en particulier sont en constante évolution. En effet, le fait que la durée de vie du matériel soit assez courte permet d'accompagner l'évolution continue des technologies ainsi que celle des demandes des expériences LHC. Ainsi depuis le démarrage du LHC, l'utilisation de la grille par l'expérience ATLAS en particulier a changé. C'est donc naturellement que les perspectives en ce qui concerne la grille de calcul et le site de grille du LPSC ont été présentées au fil du texte dans la partie II de ce document. Je ne reviendrai donc pas dessus ici mais je développerai plutôt mon projet de recherche qui concerne l'analyse des données du LHC.

9 Recherche de physique au delà du modèle standard avec le quark top au LHC

Avec des collisions proton-proton à 7, 8 puis à 13 TeV en 2015 et une très haute luminosité, le LHC est le collisionneur idéal pour rechercher des signes de physique au-delà du Modèle Standard. Par ailleurs, le quark top reste une sonde privilégiée via laquelle cette nouvelle physique pourrait se manifester. En effet, de nombreux modèles font jouer un rôle particulier à ce quark soit que sa masse et son couplage de Yukawa suggèrent qu'il joue un rôle dans la brisure de la symétrie électro-faible soit qu'ils prédisent des couplages importants entre de nouvelles particules et les fermions de la troisième famille. Des signes de nouvelle physique pourraient aussi apparaître via les mesures des propriétés du quark top et de ses désintégrations.

Afin de prolonger le travail que j'ai fait dans l'expérience DØ, j'ai choisi de rechercher des résonances se désintégrant en paires de top. Il s'agit donc de chercher des déviations dans le spectre de masse invariante $\mathrm{t\bar{t}}$ provenant de la production de paires de top par les processus du modèle standard.

Différents modèles de nouvelle physique décrivent des scénarios dans lesquels des particules lourdes se désintègrent en paires de top-antitop. Au vu des résultats récents du Tevatron sur la mesure d'asymétrie avant-arrière de production de paire de tops [60], ces modèles semblent d'autant plus intéressants à étudier. Parmi ceux qui produisent des paires de top à haute masse, on trouve les modèles de technicouleur (topcolor-assisted) avec des particules de type Z' et des modèles de dimensions supplémentaires de type Randall-Sundrum avec des gluons ou des gravitons de Kalusa-Klein. La découverte au LHC d'une particule avec les propriétés d'un Higgs standard rend le scénario de technicouleur peu probable. Cependant les analyses se veulent indépendantes des modèles et de tels scénarios peuvent rester comme référence en fonction des caractéristiques des particules qu'ils décrivent. On choisit en général comme référence une résonance étroite vis à vis de la résolution du détecteur et une résonance plus large.

Ces recherches consistent donc à sélectionner des évènements comprenant deux quarks top. On retrouve les états finaux décrits dans la mesure de la section efficace $t\bar{t}$: états finaux dileptoniques, semileptoniques ou hadroniques. Il s'agit de reconstruire la masse invariante des deux tops et d'évaluer les éventuels bruits de fond qui sont similaires à ceux décrits dans la mesure de la section efficace. Ensuite, d'éventuelles déviations par rapport aux prédictions du Modèle Standard sont recherchées dans le spectre de masse invariante $t\bar{t}$. Si aucune déviation n'est trouvée, il est alors possible de déterminer des limites en masse ou en section efficace sur les résonances recherchées.

On notera que les limites actuelles sur les masses de ces nouvelles particules sont proches du TeV. Dans cette zone en masse, les tops produits sont très boostés ce qui donne un état final particulier où les produits de désintégration des tops tendent à être très proches les uns des autres. Ainsi dans le cas d'une désintégration leptonique du top, le lepton n'est plus autant isolé que dans le cas étudié pour la mesure de section efficace dans ce document. De la même façon, pour la désintégration hadronique, les 3 jets provenant du top tendent à fusionner. Le top peut alors être reconstruit comme un seul jet de large rayon (R=1 pour ATLAS). Pour différencier l'origine de ces jets, de nouveaux outils peuvent alors être utilisés, basés sur la sous-structure des jets. Cette topologie n'est pas réservée à cette analyse et de telles outils peuvent être mis à profit dans bien d'autres études. Cet aspect de l'analyse la rend d'autant plus intéressante expérimentalement.

Cette analyse est en cours avec les données d'ATLAS prises pendant le premier run du LHC à 7 et 8 TeV. Elle se poursuivra bien sûr après le premier long arrêt du LHC qui nous donnera le temps de travailler sur les variables qui caractérisent les sous-structures des jets et qui permettent ainsi de différencier ceux qui proviennent de la désintégration de quark top. Le redémarrage du LHC à 13 TeV ouvrira alors une nouvelle fenêtre de recherche de particules trop massives ou trop peu fréquemment produites pour avoir pu être observées auparavant.

Cinquième partie

Communication vers le grand public

La communication autour des recherches que nous menons est un aspect important de nos activités. Elle passe évidemment par la publication de nos résultats dans des revues scientifiques mais nous nous devons aussi de partager avec le grand public quelles sont les problématiques que nous cherchons à résoudre, quelles sont les réponses que nous apportons et aussi la façon dont nous nous y prenons pour faire avancer nos connaissances. Nos démarches, nos outils, nos motivations sont aussi des éléments sur lesquels il est intéressant de communiquer.

Cette communication peut prendre de nombreuses formes et s'adresser à des publics très différents. Toute la difficulté se trouve là : dans la capacité à adapter notre message aux personnes à qui on s'adresse pour que nos propos leur apparaissent clairs sans toutefois être ni dévoyés ni caricaturés.

Je me suis investie depuis plusieurs années dans les différentes actions de communication du laboratoire et dans celles liées au LHC. En ce qui concerne le laboratoire, les actions de communication s'articulent autour de son service communication et documentation (3 personnes) et d'un comité de communication auquel je participe, qui regroupe, avec les personnes du service, des représentants des différents groupes et services du LPSC. Pour les évènements LHC, nous bénéficions d'un relai local intéressant avec le service de communication de la délégation régionale du CNRS ainsi que des échanges et des collaborations fructueux au sein du groupe de communication LHC de l'IN2P3, dont je fais partie aussi.

Je tire donc profit de ce document pour décrire les supports qui ont été développés au laboratoire et les principaux évènements que j'ai organisés en particulier autour de la physique du LHC.

10 Les supports

Différents supports servent à la communication générale du laboratoire. Parmi ceux-ci on notera le site web, le rapport d'activité statutaire mais aussi une nouvelle plaquette en cours de finalisation décrivant les différentes thématiques de recherche du laboratoire et ses compétences techniques. D'autres supports plus légers sont aussi développés comme l'élaboration de marque-pages par exemple.

Par ailleurs des supports ont été développés au laboratoire au fil du temps qui sont précieux pour communiquer avec le public sur nos thématiques. Je ne décrirais ici que ceux qui nous ont été utiles pour des évènements liés à la physique des particules et au LHC en particulier.

Plusieurs posters ont été conçus pour introduire nos thématiques : un poster résumant les particules et leurs interactions (figure 29), un poster sur l'historique des particules, un poster sur le LHC et ses détecteurs ... D'autre part nous possédons un exemplaire de l'exposition « Nom de code : LHC La machine à remonter le temps » , conçue juste avant le début de la prise de données du LHC qui retrace les enjeux du LHC, décrit le collisionneur et ses détecteurs ainsi que la contribution des laboratoires français. Plus ludique, des étudiants du LPSC avec l'aide des services du laboratoire, ont conçu un jeu, « l'échelle des grandeurs »

(figure 30), qui permet d'associer des longueurs (présentées en puissance de dix) à des objets allant du quark au super amas de galaxie. À l'occasion de la découverte d'une particule compatible avec le boson de Higgs, cette idée a été déclinée en une « échelle des masses ».

Un support particulièrement intéressant pour introduire le LHC et ses expériences est le "train des particules" (voir figure 31). Il s'agit d'un grand support avec une photo de la campagne Genevoise en fond sur lequel les rails d'un train électrique figurent le LHC. Deux locomotives tournant en sens inverse représentent les protons du faisceau. À chaque croisement ou « collision » des deux trains, caché aux spectateurs par un tunnel représentant le détecteur ATLAS, se déclenche l'apparition de l'affichage d'une collision simulée sur un écran. Plusieurs type d'évènements (W, top, Z' et Higgs) peuvent apparaitre de façon aléatoire mais avec des probabilités réglables. Ce « petit train » a de nombreuses vertus : en effet outre son attractivité envers tous les publics (et pas seulement les plus jeunes) il permet de parler de l'accélérateur, des collisions et de leur détection. L'affichage des évènements permet d'illustrer leurs différentes caractéristiques ce qui conduit à introduire nos techniques d'analyse. Il est aussi possible d'introduire le fait qu'on ne peut savoir à l'avance quelle collision on va observer. L'idée originale du train des particules a été dupliquée dans deux autres laboratoires de l'IN2P3.

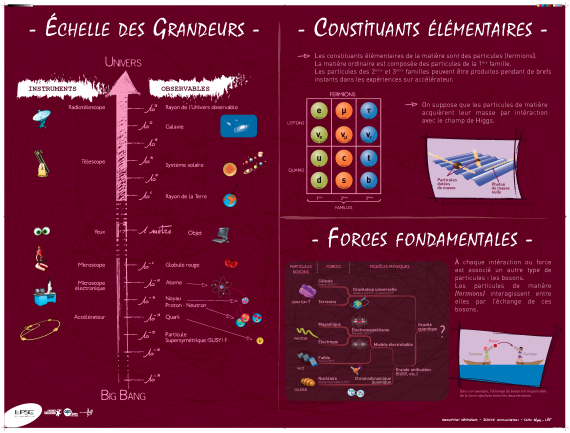


FIGURE 29 – Poster réalisé au LPSC introduisant les particules élémentaires et leurs interactions.

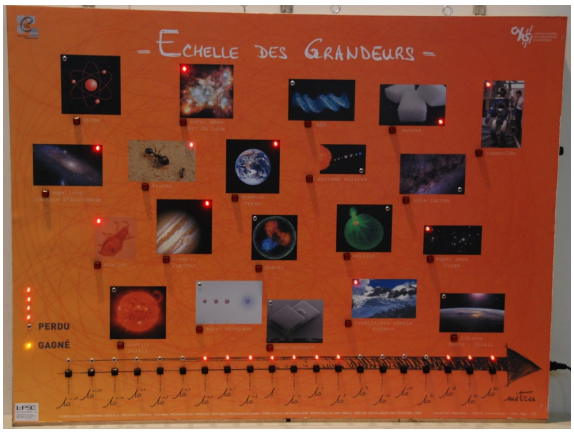


FIGURE 30 – Échelle des grandeurs : jeu permettant d'associer des longueurs (notées en puissance de 10) à des objets.

On notera aussi que le LPSC possède deux détecteurs qui permettent de visualiser des particules subatomiques : une chambre à brouillard et une chambre à étincelles (figure 32). La première permet une visualisation des particules issues de la radioactivité naturelle et la seconde permet de montrer la composante en muons du rayonnement cosmique. Ces deux détecteurs rendent plus concret le concept de particules subatomiques et sont des outils très pédagogiques pour expliquer la façon dont on détecte de tels objets inaccessibles aux sens humains.

Enfin, les supports les plus intéressants sont bien sûr les détecteurs ou outils qui sont conçus, développés, construits ou testés dans nos murs. Malheureusement, étant destinés à être installés auprès d'expériences au CERN ou ailleurs dans le monde, les possibilités de montrer des vrais éléments de ces expériences sont assez peu fréquentes. Nous avons à notre disposition des parties d'anciens détecteurs ou des prototypes (pré-échantillonneur du calorimètre d'ATLAS par exemple) en exposition. Nous avons aussi la possibilité de faire

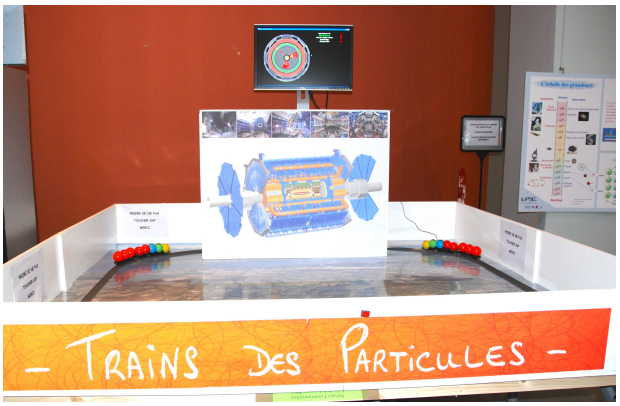


FIGURE 31 – Le train des particules

FIGURE 32 – La chambre à étincelles du LPSC.

visiter notre T2 de la grille de calcul et son refroidissement par free-cooling et nous avons eu la chance ces dernières années de pouvoir faire découvrir le hall de montage et de tests des grands modules du calorimètre électromagnétique de l'expérience ALICE, ce qui ne sera plus possible à partir de l'année prochaine.

11 Les évènements et les actions de communication

Les actions de communication peuvent prendre de nombreuses formes en fonction du public auquel elles s'adressent et des évènements autour desquels elles s'organisent.

Les scolaires représentent l'un des publics avec lequel il est gratifiant d'échanger. Nos supports n'étant pas adaptés à des enfants très jeunes, les actions que j'ai menées se sont plutôt adressées à des collégiens ou des lycéens. Pour ce public, l'exposition de l'IN2P3, qui se déplace et s'installe facilement a permis de toucher plusieurs lycées et collèges. Des séances alliant une conférence sur la physique des particules, le LHC et ses détecteurs et des discussions sur le métier de chercheur complètent la mise en place de l'exposition qui reste sur place pendant quelques semaines. Les expositions au sein du laboratoire ou à l'extérieur et les visites du laboratoire sont aussi l'occasion d'accueillir des groupes de scolaires ou d'étudiants. Collégiens et lycéens sont par ailleurs accueillis au laboratoire à l'occasion de stages de découverte et des « master class » en lien avec le CERN que des collègues organisent depuis plusieurs années. Ces actions leur permettent de mieux appréhender les métiers de la recherche en favorisant un contact avec tous les personnels du laboratoire sur plusieurs journées pour les premiers et en leur faisant vivre une journée de chercheur pour les seconds en leur permettant de mener eux-mêmes une mini-analyse des données du LHC et d'échanger leurs résultats avec d'autres classes à l'étranger.

En ce qui concerne les enfants plus jeunes, nous réfléchissons actuellement à développer des idées qui s'adresserait aux enfants de primaire pour lesquels nous n'avons pas de support adapté.

Les autres actions menées sont liées soit à des évènements récurrents soit à l'actualité

de la recherche. Chaque année par exemple, nous participons à la fête de la science. À cette occasion, le laboratoire s'ouvre aux visites pour les scolaires et le grand public ou alors nous rejoignons les lieux d'expositions de la place grenobloise. La première possibilité permet de construire des parcours assez longs (typiquement 1h30) qui donnent la possibilité de prendre le temps d'expliquer la physique, de montrer nos différentes animations et nos détecteurs et de répondre en détail aux questions d'un public souvent très motivé et intéressé. La deuxième demande des supports plus légers mais permet de toucher des personnes qui ne se déplaceraient pas spontanément jusqu'à nous. Ce type de rencontre exige une plus grande adaptabilité du point de vue du niveau du discours et du temps que les personnes sont prêtes à nous consacrer.

D'autres actions ont été déclenchées par la riche actualité de ces dernières années : démarrage du LHC, premiers résultats et découverte d'une nouvelle particule dans la recherche du boson de Higgs. Ces différentes nouvelles entrainent des demandes spontanées de conférences par différentes associations et organisations. Nous avons aussi relayés au laboratoire les principaux évènements organisés par le CERN à ces occasions pour nos collègues, le grand public et la presse. Nous avons profité de cette actualité pour organiser des expositions utilisant nos supports, avec en général la possibilité d'échanger avec les chercheurs, techniciens et ingénieurs du LPSC contribuant aux expériences ATLAS et ALICE en particulier.

Ces évènements ont été aussi l'occasion de solliciter la presse ou d'être sollicitée par elle : ainsi nos résultats scientifiques et nos différentes actions ont été relayés par la presse écrite et les télévisions locales et régionales. Notons que les formations pour apprendre à parler aux média organisées par le CERN et l'IN2P3 ont été très utiles pour aborder ces exercices avec plus de sérénité.

Ces différentes opérations, illustrées par les photos de la figure 33, n'auraient pas été possible sans le soutien du laboratoire et sans l'implication d'un grand nombre de ses personnels qui participent à ces actions souvent bénévolement et en dehors de leur temps de travail. Les retours souvent chaleureux et reconnaissants du public sont très gratifiants et préparer ces évènements avec des collègues enthousiastes est aussi un réel plaisir.

FIGURE 33 – Photos de rencontre avec le public et la presse à l'occasion d'une exposition, de la fête de la science et de la découverte d'un boson dans la recherche du Higgs.

Conclusion

La recherche en physique des particules auprès des collisionneurs nécessite la collaboration de centaines ou de milliers de personnes pour concevoir et construire des accélérateurs et des détecteurs de particules, en assurer le fonctionnement puis traiter et analyser les données qu'ils enregistrent. Chaque physicien a donc un rôle à jouer dans une ou plusieurs de ces étapes. J'ai choisi d'illustrer dans ce document deux d'entre elles auxquelles j'ai participé ces dernières années : le traitement des données avec la description de la grille de calcul et de stockage d'ATLAS et du nœud de grille du LPSC et l'analyse des données avec la mesure de la section efficace $\mathrm{t\bar{t}}$ avec les premières données du Run 2 du Tevatron. J'ai complété cet exposé par un court chapitre sur la communication au grand public qui permet de partager avec nos concitoyens nos recherches et leurs aboutissements.

Ce retour dans mes activités passées m'a permis de mesurer les progrès continus que notre communauté a faits dans la compréhension des données et les techniques d'analyse entre le démarrage de la deuxième phase de prise de données au Tevatron et les études effectuées actuellement au LHC. Je citerais par exemple la meilleure description des bruits de fond avec des jets associés (W+jets ...) et le développement des outils statistiques dont l'utilisation est devenue banale auprès des expériences LHC. L'expérience engrangée au Tevatron nous permet ainsi d'aller plus vite et plus loin dans la compréhension des données issues des collisions du LHC.

La découverte d'un boson compatible avec le boson de Higgs standard au LHC illustre combien le collisionneur, les détecteurs et les outils de traitement et d'analyse des données sont performants. La porte reste ouverte à d'autres découvertes où le quark top pourrait jouer un rôle important. Je suis heureuse de pouvoir apporter ma contribution à ce travail dans la collaboration ATLAS et accompagner de jeunes collègues dans cette aventure exigeante. Je souhaite donc continuer mes activités liées au calcul distribué et m'investir plus dans les années qui viennent dans l'analyse des données. J'ai choisi, comme je l'ai décrit dans la partie prospective de ce document, de poursuivre mes études autour du quark top en participant au développement de l'analyse sur la recherche de résonances $\mathrm{t\bar{t}}$.

Références

[1] http ://www.geant.net/Network/Pages/default.aspx

[2] F Donno et al.,*Storage resource manager version 2.2 : design, implementation, and testing experience* 2008 J. Phys. : Conf. Ser. 119 062028
http ://sdm.lbl.gov/srm-wg

[3] http ://www.egi.eu/

[4] https ://www.opensciencegrid.org

[5] http ://www.nordugrid.org/about.html

[6] http ://lcg.web.cern.ch/LCG/
LHC Computing Grid Technical Design Report, https ://espace.cern.ch/WLCG-document-repository/Technical_Documents/TDR/LCG_TDR_v1_04.pdf

[7] http ://gstat-wlcg.cern.ch/apps/pledges/summary/
http ://gstat2.grid.sinica.edu.tw/gstat/summary/WLCG_TIER/ALL/

[8] CERN-C-RRB-2005-01

[9] https ://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

[10] D.Adams et al. on behalf of the ATLAS Collaboration, *The ATLAS computing model*, CERN ATL-SOFT-2004-007, CERN-LHCC-2004-037/G-085

[11] Markus Elsing, Luc Goossens, Armin Nairz, Guido Negri, *The ATLAS Tier-0 : Overview and operational experience*, 2010 J. Phys. : Conf. Ser. 219 072011

[12] https ://twiki.cern.ch/twiki/bin/viewauth/Atlas/ExpressStream

[13] *Managing ATLAS data on a petabyte-scale with DQ2* 2008 J. Phys. : Conf. Ser. 119 062017 (http ://iopscience.iop.org/1742-6596/119/6/062017)

[14] ATLAS collaboration, *Managing ATLAS data on a petabyte-scale with DQ2*, Journal of Physics : Conference Series 119 (2008) 062017

[15] http ://rucio.cern.ch/

[16] Pool Of persistent Objects for LHC - A persistency framework, http ://lcgapp.cern.ch/project/persist

[17] http ://egee-jra1-dm.web.cern.ch/egee-jra1-dm/FTS/

[18] http ://root.cern.ch/
Brun R and Rademakers F 1997 Nucl. Instrum. Meth. A389 81 ?86

[19] T. Maeno, K. De, S. Panitkin for the ATLAS collaboration, *PD2P : PanDA Dynamic data Placement for ATLAS*, J. Phys. : Conf. Ser. 396 032070, 2012

[20] Graeme A Stewart, Vincent Garonne, Mario Lassnig, Angelos Molfetas, Martin Barisits, Donal Zhang, Ivan Calvet, Thomas Beermann, Fernando Barreiro Megino, Andrii Tykhonov, Simone Campana, Cedric Serfon, Danila Oleynik and Artem Petrosyan (for the ATLAS Collaboration), *Advances in service and operations for ATLAS data management*, J. Phys. : Conf. Ser. 368 012005, 2012
Maeno T, De K, Wenaus T, Nilsson P, Stewart G A, Walker R, Stradling A, Caballero J, Potekhin M, Smith D and Collaboration, *Overview of ATLAS PanDA Workload Management*, J. Phys. : Conference Ser. 331 072024, 2011

Molfetas A, Megino F B, Tykhonov A, Lassnig M, Garonne V, Barisits M, Campana S, Dimitrov G, Jezequel S, Ueda I and Viegas F, *Popularity framework to process dataset traces and its application on dynamic replica reduction in the ATLAS experiment*, J. Phys. : Conference Ser. 331 062018, 2011

[21] Tadashi Maeno on behalf of PanDA team and ATLAS collaboration *PanDA : distributed production and distributed analysis system for ATLAS* 2008 J. Phys. : Conf. Ser. 119 062036.
T Maeno et al, *Overview of ATLAS PanDA Workload Management*, 2011 J. Phys. : Conf. Ser. 331 072024

[22] J.T. Moscicki et al., , *Ganga : a tool for computational-task management and easy access to Grid resources"*, Computer Physics Communications, Volume 180, Issue 11, (2009), doi :10.1016/j.cpc.2009.06.016
Ganga Project Webpage : http ://cern.ch/ganga
Johannes Elmsheuser1 et al, *Reinforcing user data analysis with Ganga in the LHC era : scalability, monitoring and user-support*, 2011 J. Phys. : Conf. Ser. 331 072011

[23] P. Nilsson, J. Caballero, K. De, T. Maeno, A. Stradling , T. Wenaus for the ATLAS Collaboration *The ATLAS PanDA Pilot in Operation*, Journal of Physics : Conference Series 331 (2011) 062040

[24] Caballero, J ; Hover, J ; Love, P ; Stewart, G, *AutoPyFactory : A Scalable Flexible Pilot Factory Implementation*, ATL-SOFT-PROC-2012-045, 22 May 2012

[25] *The Condor Project*, http ://research.cs.wisc.edu/condor/description.html

[26] http ://xrootd.slac.stanford.edu/

[27] https ://twiki.cern.ch/twiki/bin/viewauth/Atlas/ConditionsDB

[28] http ://ami.in2p3.fr/opencms/opencms/AMI/www/Presentations/
Solveig Albrand, Jérôme Fulachier and Fabian Lambert, *The ATLAS metadata interface*, 2010, J. Phys. : Conf. Ser. 219 042030

[29] Jaroslava Schovancová for the ATLAS collaboration, *ATLAS Distributed Computing Monitoring tools after full 2 years of LHC data taking*, J. Phys. : Conf. Ser. 396 032095, 2012

[30] RWL Jones and D Barberis, *The Evolution of the ATLAS Computing Model* Journal of Physics : Conference Series 219 (2010) 072037 (CHEP09)

[31] http ://www.renater.fr/spip.php ?page=sommaire

[32] http ://lhcone.net/

[33] https ://twiki.cern.ch/twiki/bin/view/Atlas/CernVMFS

[34] https ://twiki.cern.ch/twiki/bin/view/Frontier/WebHome

[35] http ://eu-datagrid.web.cern.ch/eu-datagrid/

[36] B. Boutherin, *Free-cooling au LPSC* (2011) LPSC - Laboratoire de Physique Subatomique et de Cosmologie, http ://hal.in2p3.fr/in2p3-00820939

[37] https ://trac.lal.in2p3.fr/Quattor/wiki/Web ?redirectedfrom=Web/Overview

[38] http ://www.nagios.org/

[39] 2012 Review of Particle Physics : Beringer et al. (Particle Data Group), Phys. Rev. D86, 010001 (2012)

[40] CDF Collaboration, Phys. Rev. Lett. 74, 2626 (1995)
DØ Collaboration, Phys. Rev. Lett. 74, 2632 (1995)

[41] DØ Collaboration, V. Abazov *et al.*, Nucl. Instrum. Meth. A **565**, 463 (2006)
T. LeCompte and H.T. Diehl, "The CDF and DØ Upgrades for Run II", Ann. Rev. Nucl. Part. Sci. **50**, 71 (2000)

[42] DØ Collaboration, S. Abachi *et al.*, Nucl. Instrum. Methods Phys. Res. A **338**, 185 (1994)

[43] V. Abazov *et al.*, *The Muon System of the Run II Detector*, FERMILAB-PUB-05-034-E

[44] N. Kidonakis, R. Vogt, "Next-to-Next-to-leading Order Soft-Gluon Corrections in Top Quark Hadroproduction", Phys. Rev. D 68, 114014 (2003)

[45] *Review of Particle Physics*, S. Eidelman et al., Phys. Lett. B **592**, 1 (2004)

[46] F. Chevallier, S. Crépé-Renaudin, *Measurement of the* $\mathbf{t\bar{t}}$ *Production Cross Section at* $\sqrt{\mathbf{s}} = \mathbf{1.96}$ *TeV in the Lepton+Jets Final State using Soft Muon Tagging on the first* $\mathbf{425\ pb^{-1}}$ *of DØ Run II data*, DØ Note 5115

[47] F. Chevallier, *Mesure de la section efficace de production de quarks top en paires dans le canal lepton+jets à DØ et à ATLAS et interprétation en terme de boson de Higgs chargé dans ATLAS.*, 10 mai 2007, thèse de doctorat

[48] DØ collaboration, *Measurement of the* $\mathbf{t\bar{t}}$ *Production Cross Section in* $\mathbf{p\bar{p}}$ *Collision at* $\sqrt{\mathbf{s}} = \mathbf{1.96}$ *TeV Using Soft Muon b-tagged Lepton+Jets Events*, DØ Note 5257-Conf

[49] http ://www-d0.fnal.gov/computing/algorithms/# intro

[50] Aran Garcia-Bellido, Sara Lager, Flera Rizatdinova, Ariel Schwartzman and Gordon Watts, *Primary Vertex certification in p14*, DØ Note 4320, Apr 2004

[51] Blazey, G. et al., *Run II Jet Physics*, DØ Note 3750, April 2000

[52] J. Kozminski et al., *The electron likelihood in p14*, DØ Note 4449, November 2003 ;
Ashish Kumar, Brajesh Choudhary, Joseph Kozminski, Robert Kehoe, Jon Hays, Jan Stark, *Electron Likelihood Study*, DØ Note 4769, March 2005

[53] K. Hanagaki and J. Kasper, *Identification of b-jet by Soft Muon*, DØ note 4867, August 15, 2005

[54] Mangano, Michelangelo L. and Moretti, Mauro and Piccinini, Fulvio and Pittau, Roberto and Polosa, Antonio D., ALPGEN, *a generator for hard multiparton processes in hadronic collisions*, JHEP **07** 001 (2003), hep-ph/0206293

[55] http ://mlm.home.cern.ch/mlm/alpgen/

[56] T. Andeen et. al., FERMILAB-TM-2365

[57] Nikola Makovec and Jean-Francois Grivaz Shifting, Smearing and Removing Simulated Jets D0 Note Number :004914 Date : 8/31/05

[58] DØ Collaboration, *Measurement of the* $\mathrm{t\bar{t}}$ *production cross section in pp collision at* $\sqrt{s} = 1.96\ TeV$ *using kinematic characteristics of lepton+jets events*, PRD 76, 092007 (2007)

[59] DØ Collaboration, *Measurement of the* $\mathrm{t\bar{t}}$ *production cross section in pp collision at* $\sqrt{s} = 1.96\ TeV$ *using secondary vertex tagging*, PRD 74, 112004 (2006)

[60] CDF Collaboration, T. Aaltonen et al., *Evidence for a Mass Dependent Forward-Backward 864 Asymmetry in Top Quark Pair Production*, Phys.Rev. D83 (2011) 112003, 865 arXiv :1101.0034 [hep-ex].

Annexes

A Calcul de la relation entre les variables de la fonction de vraisemblance utilisée pour la mesure de section efficace $t\bar{t}$

L'équation (12) peut être développée en fonction des variables M_j. Pour simplifier la notation, le calcul est fait pour un canal donné ce qui permet de laisser de côté l'indice j :

$$\begin{aligned}
\mathrm{N}^{\mathrm{tag}}_{\mathrm{stricte}} &= \mathrm{N}^{\mathrm{tag}}_{\mathrm{t\bar{t},\ stricte}} + \mathrm{N}^{\mathrm{tag}}_{\mathrm{bruit\ de\ fond,\ stricte}} \\
\mathrm{N}^{\geq 1}_{\mathrm{stricte}} &= \mathrm{N}^{\mathrm{tag}}_{\mathrm{t\bar{t},\ stricte}} + \mathrm{N}^{\mathrm{tag}}_{\mathrm{MCbdf,\ stricte}} + \mathrm{N}^{\mathrm{tag}}_{\mathrm{W+jets,\ stricte}} + \mathrm{N}^{\mathrm{tag}}_{\mathrm{multijets,\ stricte}} \\
\mathrm{N}^{\geq 1}_{\mathrm{stricte}} &= \mathrm{P}^{\mathrm{tag}}_{\mathrm{t\bar{t}}} \,.\, \mathrm{N}^{\mathrm{sel}}_{\mathrm{t\bar{t},\ stricte}} + \mathrm{P}^{\mathrm{tag}}_{\mathrm{MCbdf}} \,.\, \mathrm{N}^{\mathrm{sel}}_{\mathrm{MCbdf,\ stricte}} + \mathrm{P}^{\mathrm{tag}}_{\mathrm{W+jets}} \,.\, \mathrm{N}^{\mathrm{sel}}_{\mathrm{W+jets,\ stricte}} + \mathrm{N}^{\mathrm{tag}}_{\mathrm{multijets,\ stricte}}
\end{aligned}$$

sel et tag font référence respectivement aux évènements ayant passé la présélection et l'étiquetage des b ; $\mathrm{P}^{\mathrm{tag}}$ correspond à la probabilité d'étiquetage que possèdent ces évènements. MCbdf regroupe les évènements de bruit de fond évalués avec le Monte Carlo (diboson, single top, Z+jets) et $\mathrm{t\bar{t}}$ les évènements du canal lepton+jets et du canal dilepton sélectionnés. On suppose ici que ces deux canaux ont la même section efficace.

$\mathrm{N}^{\mathrm{sel}}_{\mathrm{W+jets,\ stricte}}$, le nombre d'évènements W+jets, est obtenu en soustrayant aux évènements avec un vrai lepton isolé $\mathrm{N}^{\mathrm{sel}}_{\mathrm{vrai\ lepton,\ stricte}}$ les évènements $\mathrm{t\bar{t}}$, $\mathrm{N}^{\mathrm{sel}}_{\mathrm{t\bar{t},\ stricte}}$, et les bruits de fond contenant un vrai lepton $\mathrm{N}^{\mathrm{sel}}_{\mathrm{MCbkd,\ stricte}}$:

$$\mathrm{N}^{\mathrm{sel}}_{\mathrm{W+jets,\ stricte}} = \mathrm{N}^{\mathrm{sel}}_{\mathrm{vrai\ lepton,\ stricte}} - \mathrm{N}^{\mathrm{sel}}_{\mathrm{t\bar{t},\ stricte}} - \mathrm{N}^{\mathrm{sel}}_{\mathrm{MCbdf,\ stricte}} \tag{17}$$

L'équation (17) devient donc :

$$\begin{aligned}
\mathrm{N}^{\geq 1}_{\mathrm{stricte}} &= \mathrm{P}^{\mathrm{tag}}_{\mathrm{W+jets}} \,.\, \mathrm{N}^{\mathrm{sel}}_{\mathrm{vrai\ lepton,\ stricte}} + \mathrm{N}^{\mathrm{tag}}_{\mathrm{QCD,\ stricte}} && (18) \\
&\quad + (\mathrm{P}^{\mathrm{tag}}_{\mathrm{t\bar{t}}} - \mathrm{P}^{\mathrm{tag}}_{\mathrm{W+jets}}) \,.\, \mathrm{N}^{\mathrm{sel}}_{\mathrm{t\bar{t},\ stricte}} + (\mathrm{P}^{\mathrm{tag}}_{\mathrm{MCbkd}} - \mathrm{P}^{\mathrm{tag}}_{\mathrm{W+jets}}) \,.\, \mathrm{N}^{\mathrm{sel}}_{\mathrm{MCbkd,\ stricte}} && (19)
\end{aligned}$$

Le nombre d'évènements multi-jets est obtenu par la méthode de la matrice après étiquetage des b (5) :

$$\begin{aligned}
\mathrm{N}^{\mathrm{tag}}_{\mathrm{multijets,\ stricte}} &= \varepsilon_{\mathrm{multijets}} \frac{\varepsilon_{\mathrm{sig}}\, \mathrm{N}^{\mathrm{tag}}_{\mathrm{N_{lâche}}} - \mathrm{N}^{\mathrm{tag}}_{\mathrm{stricte}}}{\varepsilon_{\mathrm{sig}} - \varepsilon_{\mathrm{QCD}}} \\
\mathrm{N}^{\mathrm{tag}}_{\mathrm{multijets,\ stricte}} &= \varepsilon_{\mathrm{multijets}} \frac{\varepsilon_{\mathrm{sig}}\, (\mathrm{N}^{\geq 1}_{\mathrm{stricte}} + \mathrm{N}^{\geq 1}_{\mathrm{\bar{N}_{lâche}-stricte}}) - \mathrm{N}^{\geq 1}_{\mathrm{stricte}}}{\varepsilon_{\mathrm{sig}} - \varepsilon_{\mathrm{multijets}}}
\end{aligned} \tag{20}$$

Le nombre d'évènements avec un vrai lepton après la présélection stricte est obtenu via la méthode de la matrice appliquée avant étiquetage des b (6) :

$$\begin{aligned}
\mathrm{N}^{\mathrm{sel}}_{\mathrm{vrai\ lepton}} &= \varepsilon_{\mathrm{sig}} \frac{\mathrm{N}_{\mathrm{stricte}} - \varepsilon_{\mathrm{multijets}}\, \mathrm{N}_{\mathrm{N_{lâche}}}}{\varepsilon_{\mathrm{sig}} - \varepsilon_{\mathrm{multijets}}} \\
\mathrm{N}^{\mathrm{sel}}_{\mathrm{vrai\ lepton}} &= \varepsilon_{\mathrm{sig}} \frac{\mathrm{N}^{0}_{\mathrm{stricte}} + \mathrm{N}^{\geq 1}_{\mathrm{stricte}} - \varepsilon_{\mathrm{multijets}}\, (\mathrm{N}^{0}_{\mathrm{stricte}} + \mathrm{N}^{\geq 1}_{\mathrm{stricte}} + \mathrm{N}^{0}_{\mathrm{N_{lâche}-stricte}} + \mathrm{N}^{\geq 1}_{\mathrm{\bar{N}_{lâche}-stricte}})}{\varepsilon_{\mathrm{sig}} - \varepsilon_{\mathrm{multijets}}}
\end{aligned} \tag{21}$$

En introduisant les équations (20) et (21) dans l'équation (19), $N^{\geq 1}_{\text{stricte}}$ s'exprime en fonction des autres paramètres de la fonction de vraisemblance :

$$\begin{aligned} N^{\geq 1}_{\text{stricte}} &= a\, N^{0}_{\text{stricte}} - ab\, N^{0}_{N_{\text{lâche}}\,-\,\text{stricte}} + b\, N^{\geq 1}_{N_{\text{lâche}}\,-\,\text{stricte}} \\ &\quad + c\,(P^{\text{tag}}_{\text{MCbdf}} - P^{\text{tag}}_{\text{W+jets}})\, N^{\text{sel}}_{\text{MCbdf, stricte}} + c\,(P^{\text{tag}}_{t\bar{t}} - P^{\text{tag}}_{\text{W+jets}})\, N^{\text{sel}}_{t\bar{t},\,\text{stricte}}(\sigma(t\bar{t})) \end{aligned}$$

où

$$\begin{aligned} a &= \frac{P^{\text{tag}}_{\text{W+jets}}}{1 - P^{\text{tag}}_{\text{W+jets}}}, \\ b &= \frac{\varepsilon_{\text{multijets}}}{1 - \varepsilon_{\text{multijets}}}, \\ c &= \frac{\varepsilon_{\text{sig}} - \varepsilon_{\text{multijets}}}{\varepsilon_{\text{sig}} \,.\, (1 - \varepsilon_{\text{multijets}}) \,.\, (1 - P^{\text{tag}}_{\text{W+jets}})}. \end{aligned}$$

B Distributions cinématiques

La comparaison entre les données sélectionnées (après présélection stricte et étiquetage des b) et la somme des bruits de fond évalués avec le signal $t\bar{t}$ est illustré ici. Les figures 34 et 35 montrent dans les canaux e+jets et μ+jets quelques exemples de distributions cinématiques :

- impulsion transverse des leptons isolés (électron ou muon),
- énergie transverse du jet de plus grande impulsion,
- masse transverse du boson W,
- somme de l'énergie transverse des jets dans l'évènement : variable H_T.

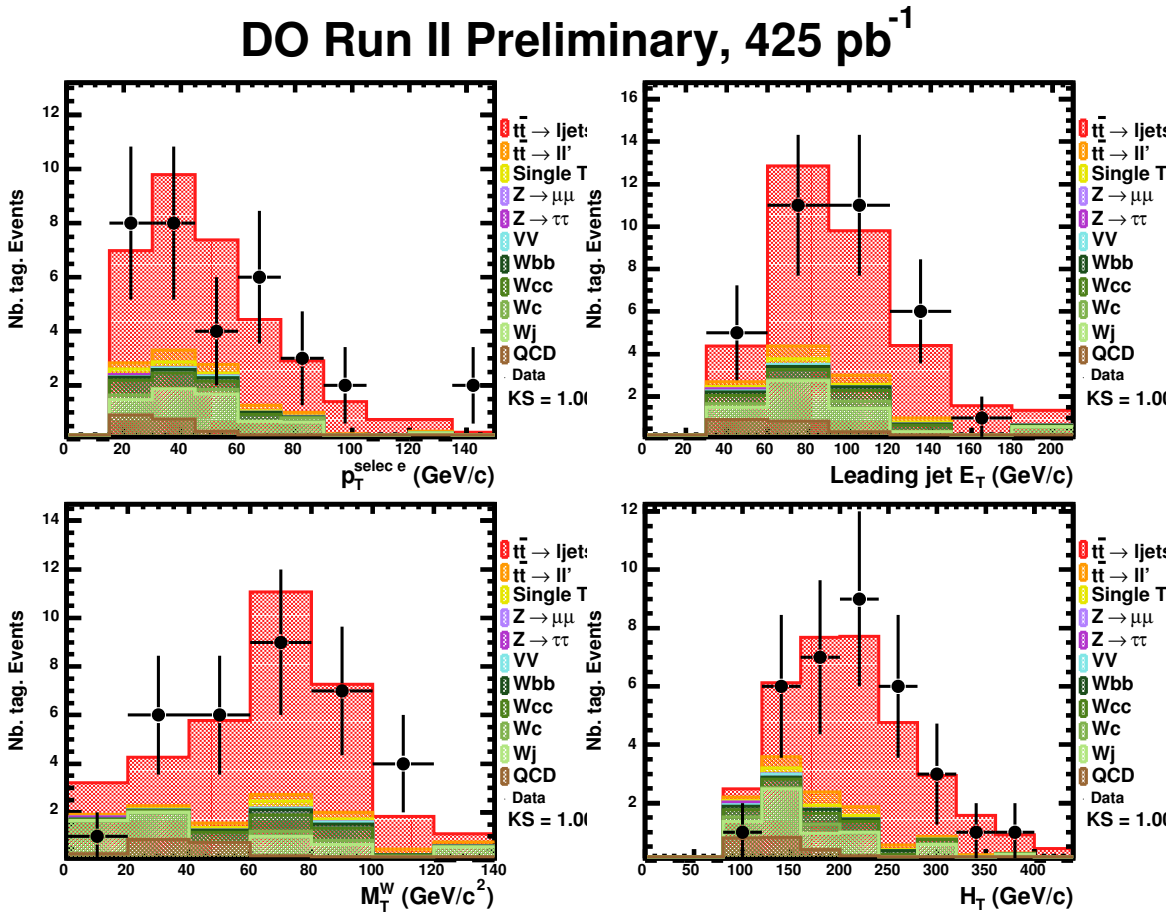


FIGURE 34 – Données comparées à l'évaluation de la somme des bruits de fond et du signal $t\bar{t}$ pour les évènements avec 3 jets ou plus dans le canal e+jets. Le signal $t\bar{t}$ est décrit en utilisant la section efficace mesurée par l'analyse.

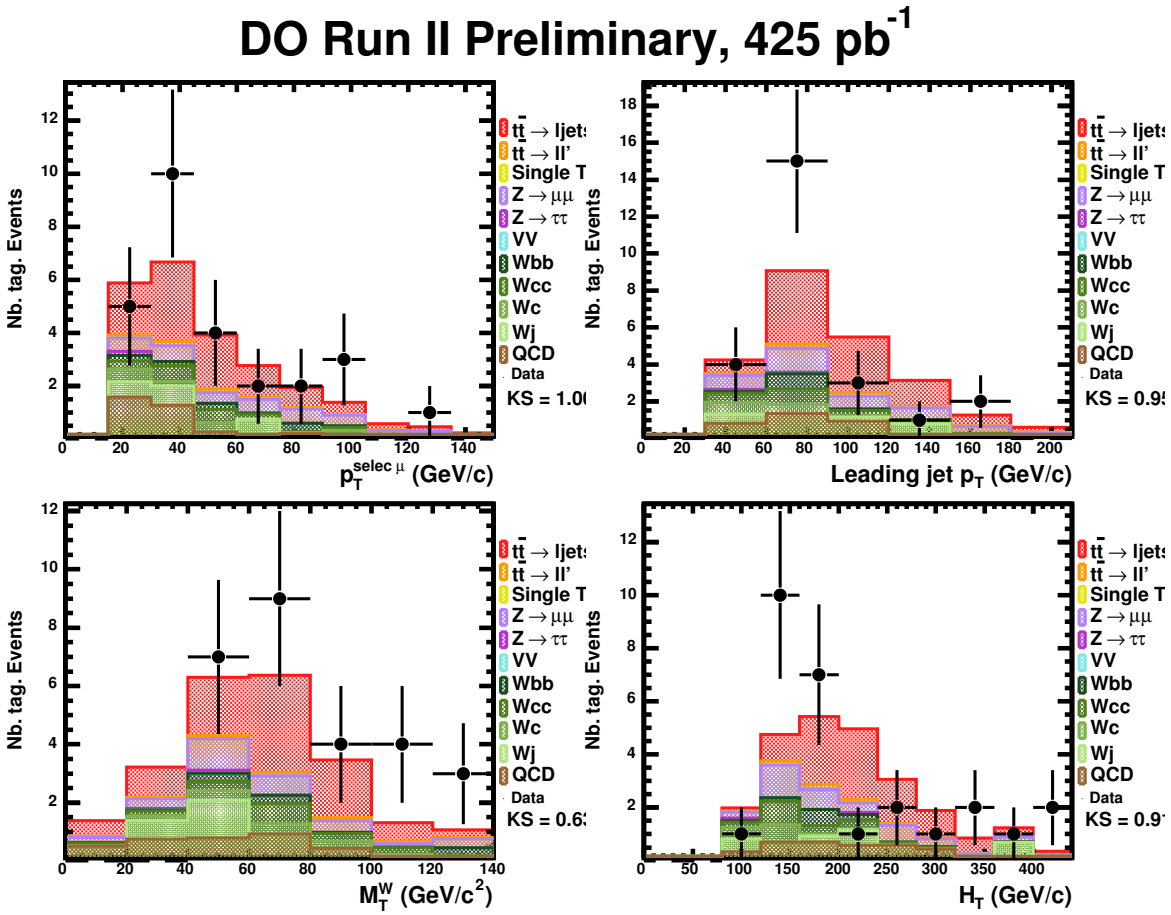


FIGURE 35 – Données comparées à l'évaluation de la somme des bruits de fond et du signal $t\bar{t}$ pour les évènements avec 3 jets ou plus dans le canal μ+jets. Le signal $t\bar{t}$ est décrit en utilisant la section efficace mesurée par l'analyse.